Multi-Instance Active Learning with Online Labeling for Object Recognition

Kimia Salmani and **Mohan Sridharan**
Department of Computer Science
Texas Tech University, USA
{kimia.salmani, mohan.sridharan}@ttu.edu

Abstract

Robots deployed in domains characterized by non-deterministic action outcomes and unforeseen changes frequently need considerable knowledge about the domain and tasks they have to perform. Humans, however, may not have the time and expertise to provide elaborate or accurate domain knowledge, and it may be difficult for robots to obtain many labeled training samples of domain objects and events. For widespread deployment, robots thus need the ability to incrementally and automatically extract relevant domain knowledge from multimodal sensor inputs, acquiring and using human feedback when such feedback is necessary and available. This paper describes a multiple-instance active learning algorithm for such incremental learning in the context of building models of relevant domain objects. We introduce the concept of *bag uncertainty*, enabling robots to identify the need for feedback, and to incrementally revise learned object models by associating visual cues extracted from images with verbal cues extracted from limited high-level human feedback. Images of indoor and outdoor scenes drawn from the IAPR TC-12 benchmark dataset are used to show that our algorithm provides better object recognition accuracy than a state of the art multiple-instance active learning algorithm.

1 Introduction

Sophisticated algorithms are enabling the use of robots[1] in application domains such as search and rescue, surveillance, and health care. In such domains characterized by non-deterministic action outcomes and unforeseen changes, it is difficult for robots to operate without considerable domain knowledge. Humans, however, may not have the expertise or time to interpret raw sensor data, or to provide accurate domain knowledge, and it may be difficult for robots to obtain many labeled training samples of domain objects and events. Widespread deployment of robots thus poses the challenge of incrementally acquiring relevant domain knowledge using multimodal cues extracted from sensor inputs and high-level human feedback based on need and availability.

Consider a room with multiple objects, with the labels of some of these objects being known. A robot entering this room to retrieve an object may not possess a model of the corresponding object class, or the specific instance in this room may be different. The robot may thus have to solicit help from humans to locate the object and obtain labeled samples to build or revise the corresponding object model. Non-expert humans can assist the robot if they are allowed to provide different types of cues, and their response can be limited to high-level feedback, e.g., confirming the presence or absence of specific objects or object classes in the room. This paper describes an algorithm for such incremental learning of object models using visual and verbal cues. We build on multiple-instance learning (MIL) algorithms that learn from labeled *bags* of *instances*. For visual recognition, each image is a "bag", while segmented salient regions of interest (ROIs) are "instances"—see Figure 1. Labeling bags instead of instances reduces human effort. Although active learning algorithms have been developed to revise such learned object models using verbal or textual feedback from humans, it is difficult to associate human feedback with specific image ROIs; existing multiple-instance active learning (MIAL) algorithms therefore acquire feedback only on instances in the labeled training set. Our algorithm overcomes this limitation through the following contributions:

Figure 1: For visual object recognition, images are represented as bags and specific instances correspond to segmented image regions of interest. Image drawn from the IAPR TC-12 benchmark dataset (Escalante et al. 2010).

- The concept of *bag uncertainty* is introduced to quantify the uncertainty about the presence of specific objects in

[1] We use "agent," "learner," and "robot" interchangeably.

an image. Object models learned from an initial set of images are used in conjunction with existing active learning algorithms to classify ROIs in new images and compute bag uncertainty.

- Human feedback, when available, is directed towards images with high (bag) uncertainty. Verbal cues from humans about bags are associated with specific image ROIs, generating (bag) labels for the new images used to revise the learned object models.

These contributions are a significant step towards incremental learning for robots collaborating with non-experts. The algorithm is evaluated by comparing it with a state of the art MIAL algorithm on images of indoor and outdoor scenes drawn from the IAPR TC-12 benchmark dataset.

2 Related Work

This section motivates our algorithm by briefly describing related work in multi-instance learning, active learning and the use of verbal cues.

2.1 Multiple-Instance Learning

Multiple-instance learning (MIL) is a supervised learning method that uses labels assigned to bags instead of instances to decrease the labeling effort. For classification tasks, positive bags for a class have at least one positive instance, while negative bags only have negative instances. Maron and Lozano-Perez (1997) introduced the Diverse Density framework for MIL, which identifies feature space points with high diverse density, and Maron and Ratan (1998) used this framework to classify natural scenes. Yang et al. (2000) applied MIL to classify images with a range of desired objects. Zhang and Golman (2001) proposed the EM-DD algorithm based on the expectation-maximization algorithm. More recently, Zha et al. (2008) developed an algorithm for multi-label MIL, while Zhang et al. (2013) enabled MIL from multiple information sources. The basic MIL algorithms do not support incremental acquisition and use of bag labels.

2.2 Active Learning

Active learning seeks to minimize the labeling effort in supervised learning. Settles (2009) provides a survey of *pool-based*, *membership query* and *stream-based* active learning algorithms. Settles, Craven and Ray (2008) introduced multiple-instance active learning (MIAL) by combining *pool-based* active learning with MIL. The learner is trained with a small set of labeled data and selectively asks queries from a larger pool of static unlabeled data. Druck et. al (2009) used pool-based active learning to solicit labels for features instead of instances, e.g., *water* and *garbage* are considered features for the label *utilities*. MIAL algorithms have been used for different classification and regression tasks (Settles 2012). Active learning has also been used for other applications that use textual, verbal and/or contextual information, e.g., Siddiquie and Gupta (2010) use active learning to learn appearance and contextual models that are used to pose label queries for multi-class classification.

2.3 Verbal Understanding

Sophisticated algorithms have enabled robots to use multimodal data, e.g., vision, speech and text (Aboutalib and Veloso 2010; Cantrell et al. 2010; Hawes et al. 2010), but learning from multimodal cues poses the challenge of associating the information extracted from each cue. Existing work has predominantly established these associations using pre-specified rules. Recently, Swaminathan et. al (2012) proposed an algorithm to automatically learn probabilistic associations between visual and verbal vocabularies for posing relevant disambiguating queries, but only reported proof of concept results with simplistic tabletop objects.

This paper addresses key limitations of existing work. Our algorithm learns associations between visual and verbal cues, and supports incremental multiple-instance active learning with previously unseen images to revise learned models of domain objects.

3 Proposed Algorithm

Figure 2 is an overview of our architecture for MIAL with online verbal labeling. A set of labeled images are used as input to an MIL algorithm, generating initial models for desired object classes. These models are used to classify ROIs in new (test) images, ranking these images based on classification uncertainty computed using a bag uncertainty measure. Human feedback is solicited when such feedback is necessary and available. Verbal (label) inputs from humans are associated with visual cues extracted from these images, generating bag labels that support the inclusion of the new images to revise the learned object models. The architecture's components thus enable the learner to: (1) establish the need for human feedback; (2) solicit human feedback on relevant images; and (3) learn from the human feedback. Specific details are provided below.

3.1 Estimate Need for Human Feedback

The first key function of our architecture is to estimate the need for human feedback. Unlike existing MIAL methods that obtain human feedback on specific instances in a static set of labeled images, our approach supports incremental (online) learning. An initial set of images with labeled ROIs are used to create positive and negative bags for the different object classes under consideration, e.g., *window, car* and *street*. Object models learned from these bags using MIL are used to classify ROIs in new (i.e., previously unseen) images. The classification uncertainty for the entire image (or bag) is computed by introducing a *bag uncertainty* measure. For a specific object class under consideration, the bag uncertainty is estimated as follows:

$$U(B_i) = 2p_i(1-p_i) \tag{1}$$

where B_i is the i^{th} bag and p_i is the probability of bag i being positive. The probability of i^{th} bag being positive is estimated by combining the support provided by each instance in the bag (Settles, Craven, and Ray 2008):

$$\begin{aligned} p_i = P(y_i = 1|B_i) &= Softmax_\alpha(p_{i1},...,p_{in}) \\ &= \frac{\sum_{j=1}^{n} p_{ij} e^{\alpha p_{ij}}}{\sum_{j=1}^{n} e^{\alpha p_{ij}}} \end{aligned} \tag{2}$$

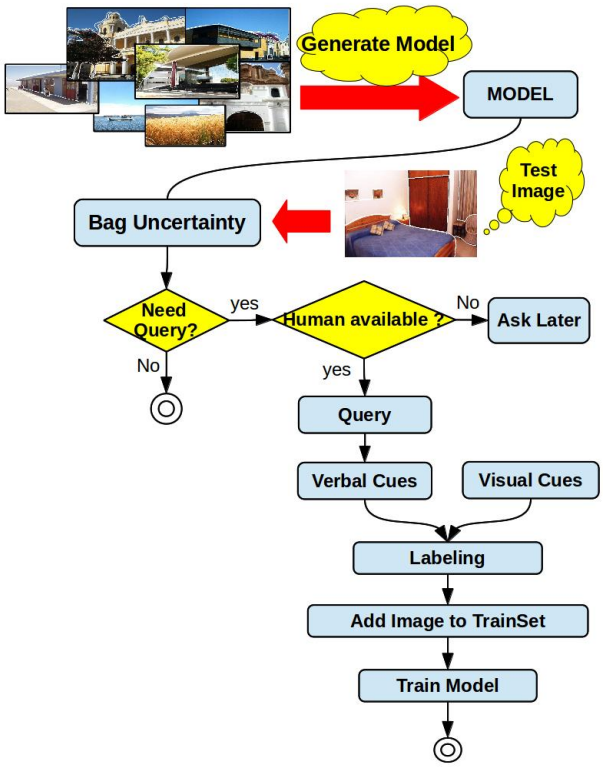


Figure 2: Architecture for multi-instance active learning with online (verbal) labeling.

where α is a non-zero constant that is set experimentally. An image with a high bag uncertainty for a specific label is a good candidate for soliciting human feedback for the corresponding object class. For each object class, new images are thus ranked based on bag uncertainty. When there are multiple such images/bags with high uncertainty, the learner establishes the need for human feedback.

3.2 Soliciting Human Feedback

The second key function of the architecture is to solicit human feedback on appropriate bags or images. Our algorithm uses a combination of *stream-based* and *pool-based* active learning to solicit human input. Each new image is considered as a candidate and feedback is solicited only on images with a suitably high bag uncertainty, while other images are not used for learning; this corresponds to stream-based active learning. At the same time, feedback is solicited only if a human is available for providing feedback. In the absence of a human willing to provide feedback, images with suitably high bag uncertainty are ranked and stored for use when a human is available; this corresponds to pool-based active learning. Human feedback is obtained by displaying the selected image and recording verbal inputs.

3.3 Learning from Human Feedback

Human feedback consists of simple sentences about the entire image or a specific ROI in the image. A sentence that contains negative expressions labels the bag and each of its

Figure 3: Illustrative example of POS tags and BIO tags for words in a sentence.

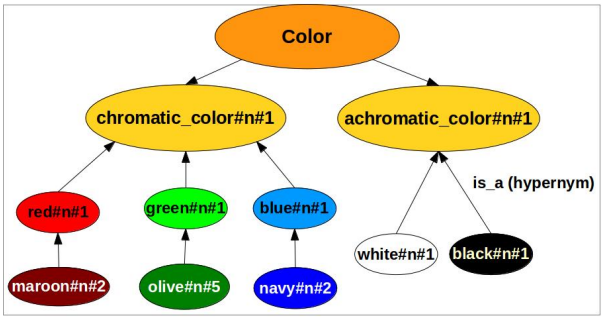


Figure 4: Illustration of hypernym relation in WordNet.

instances negative for the corresponding object class and features. A sentence with a positive expression may, in addition to object labels, provide information about specific visual features (e.g., color or size) of specific ROIs in the image under consideration. As an illustrative example of associating such verbal cues with visual cues extracted from specific ROIs in the image, consider a specific sentence and assume that human input contains color and object labels. Also assume that each color label corresponds to specific (known) RGB values. To associate the object label with a specific instance, a histogram of the RGB values of pixels is extracted from each ROI (instance) in this image (bag). These histograms are compared with the expected RGB values for the color label in the sentence. The instance corresponding to the histogram that best matches the expected RGB value is assigned the corresponding object label; the model for the corresponding object class can now be revised using this labeled instance or the image (as a positive bag).

The degree of match between the observed and expected color values can be computed using different distance measures, depending on the representation of the color values for each color label. For instance, if expected color values are triples, the Euclidean distance metric is used to identify the instance whose average RGB value best matches the triple of the color label:

$$ID^* = \arg\min_i \sqrt{(R_i - R_c)^2 + (G_i - G_c)^2 + (B_i - B_c)^2} \quad (3)$$

where R_i, G_i, B_i are the average RGB values of instance i, and R_c, G_c, B_c are RGB values of the color label in the sentence. If the color values are represented as normalized histograms, the Jenson-Shannon distance measure is used for comparison (Cover and Thomas 2006). If the human input is ambiguous (e.g., matches multiple instances), one of the valid candidates is chosen. Incorrect matches are handled by the underlying MIL algorithm.

To perform the association described above, relevant verbal cues have to be first extracted from the sentence. Indi-

vidual words in a sentence under consideration may play different roles such as *noun* and *verb*. To identify these roles, the sentence is parsed to automatically obtain Part of Speech (POS) tags and BIO tags. POS tags are extracted for each word using Stanford Log-Linear POS Tagger (Toutanova et al. 2003) with Penn Treebank tag set (Marcus, Marcinkeiwicz, and Santorini 1993). Common POS tags are *noun, adjective, verb, adverb, determiner*, which are denoted by: *NN, JJ, VBZ, RB, DT* respectively. BIO tags are extracted according to the IOB2 convention (Ratnaparkhi 1998), with *B*, *I* and *O* representing the beginning, inside and outside of the feature or property label. In the current example, two types of property labels are considered, e.g, object category (CAT) and color (COL). For example, Figure 3 considers the sentence:"There is a blue door in the image" to illustrate the POS and BIO tags. This sentence provides information about a specific ROI in the image. It states that the blue object in the image is a door. This image can thus be a positive bag for object class *door*, and the matched instance in the image may be used as a positive (labeled) instance. Similarly, the sentence may provide negative information by including words *not* or *no*.

Given the verbal tags, the semantic meaning of the relevant words (e.g., color and category labels) is extracted using WordNet, a large lexical database for nouns, verbs, adjectives, and adverbs, which are grouped into sets of cognitive synonyms (synsets) (Fellbaum 1995; 1998). Synsets consist of all possible conceptual-semantics. Each WordNet's synsets is linked to other synsets with lexical relations such as *synonyms, antonyms, hypernyms* and *hyponyms*.

Color values such as *red, green, blue* have the same hypernym *chromatic_color* and encode a *is_a* relation; red is a chromatic color. Color values such as *black, white* have the same hypernym *achromatic_color*, while the second level hypernym of some colors like *aqua, lime* is *chromatic_color*, since olive is a type of green and green is a chromatic color. Figure 4 illustrates the *is_a* relation in WordNet. The similarities are calculated between the given vocabularies and the vocabularies in a dictionary of words that can be used as verbal labels, based on the *lexical* similarity and *content*. In the current example, the dictionary consists of all color and object labels that are likely to be used. Each word is represented as: "word#pos#sense", where "*pos*" refers to the POS, with values *n* for noun, *a* for adjective, and *v* for verb; and "*sense*" refers to the word's special semantic ID.

Extracting information from visual and verbal inputs, and associating them appropriately, helps the learner make best use of the available information to incrementally revise object models, as evaluated in the next section.

4 Experimental Setup and Results

We implemented and evaluated our algorithm on images of indoor and outdoor scenes extracted from the *IAPR TC-12 Benchmark* database (Escalante et al. 2010). IAPR TC-12 has $\approx$ 20000 images in 40 folders. Figure 5 shows examples of images of indoor and outdoor scenes that were chosen from the IAPR TC-12 database. Each image is segmented into salient ROIs and annotated as shown in Figure 5; existing computer vision algorithms can be used to extract such ROIs from images. For the experiments reported below, the visual features extracted from an ROI consists of:

$$< r_i, c_i, hsb_{i1}, hsb_{i2}, hsb_{i3}, ..., hsb_{i64} >$$

where r_i represents the ratio of boundary and area of the i^{th} ROI, c_i represents the convexity, and $hsb_{i1}...hsb_{i64}$ represents the histogram in HSB color space with 64 bins.

Experimental trials used images drawn from 15 folders for training object models, and incrementally revised the models using images randomly selected from four folders. We use a state of the art MIAL algorithm as the baseline (Settles, Craven, and Ray 2008)—our algorithm and the baseline are labeled *Bag Uncertainty* (BU) and *Multiple-instance Uncertainty* (MIU) respectively in the figures below. The underlying MI learner for both algorithms starts with 200 randomly drawn positive bags and 200 randomly drawn negative bags; the model is then revised and evaluated using the remaining bags. Note that incremental learning in BU occurs with a new learning set consisting of unseen images; in MIU, learning only takes place using the instances in the positive bags of the larger training set. The MIL model uses $\alpha = 2.5$ for the Softmax function (Equation 2) and is trained by minimizing squared loss via L-BFGS (Nocedal and Wright 1999). Results are averaged over ten independent repetitions for each object class or category.

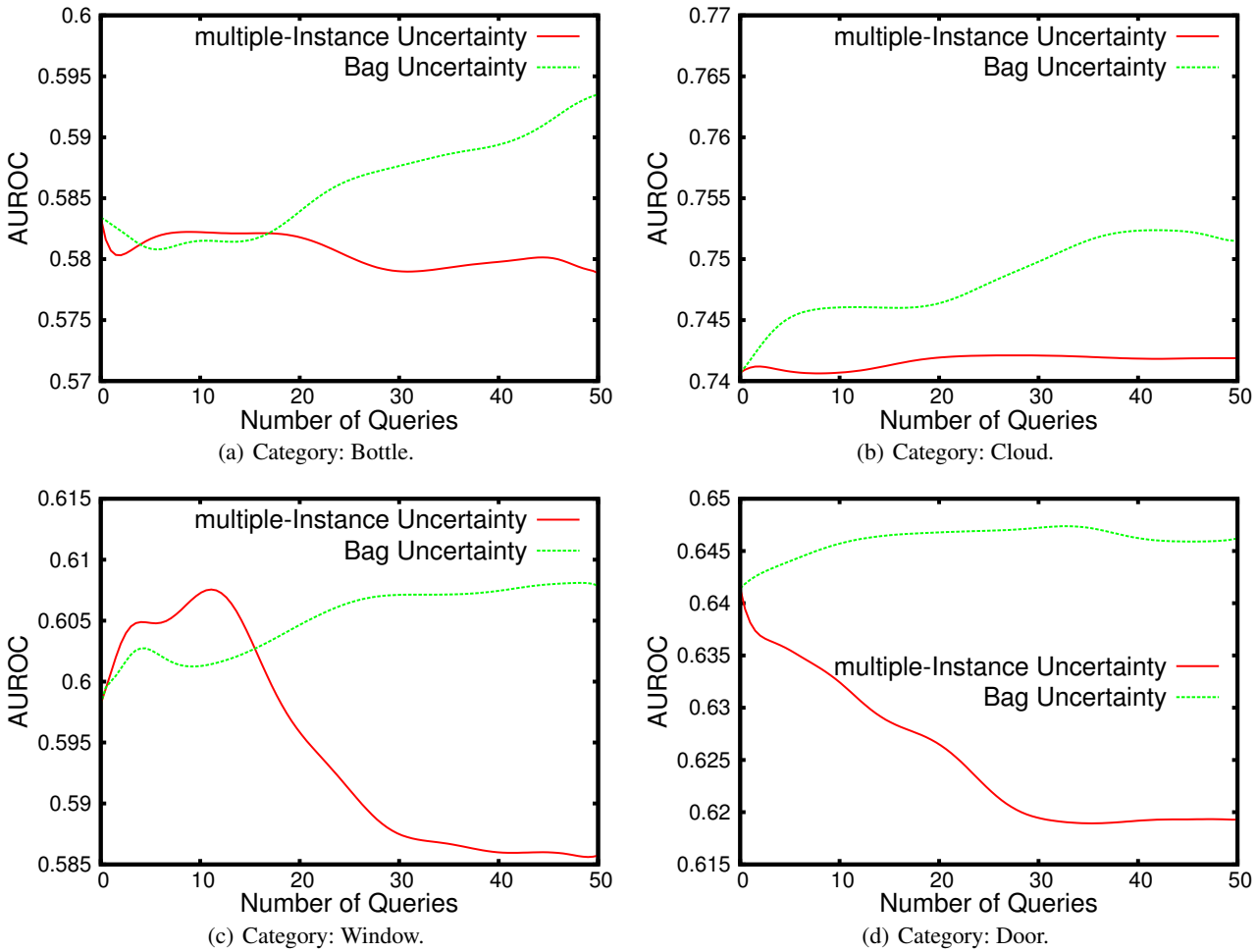
The two algorithms are compared by constructing learning curves that plot the area under the ROC curve (AUROC) as a function of the number of instances posed as queries for human feedback for each object category. The starting point in all experiments is the AUROC for a model trained from the labeled bags in the training set without including any human feedback. For our algorithm, stream-based active learning uses a threshold of 0.3, i.e., images with bag uncertainty ≥ 0.3 are valid candidates for human (verbal) input. Figure 6 shows some illustrative examples of such learning curves for four object classes. We observe that assigning labels to new images and instances, and using this information to revise the learned object models, improves the object recognition performance. A summary of performance over all the object classes included in our study is presented in Table 1; the differences between BU and MIU are statistically significant.

As discussed by Settles et. al (2008), MIL is a challenging problem, typically resulting in much lower classification accuracy that traditional object recognition algorithms. The improvements in classification accuracy provided by our algorithm thus have practical benefits.

5 Conclusions and Future Work

This paper described an algorithm that introduced the bag uncertainty measure to compute classification uncertainty in multiple-instance active learning. Unlike existing MIAL algorithms, our algorithm is able to rank new images based on classification uncertainty, and solicit human input when it is necessary and available by combining stream-based and pool-based active learning algorithms. Furthermore, verbal cues extracted from the human input are associated with the visual cues extracted from specific instances in images, enabling the learner to make best use of the available information. Incrementally revising the learned object models re-

Figure 5: Examples of images from the SAIAPR TC-12 benchmark (Escalante et al. 2010).



(a) Category: Bottle.

(b) Category: Cloud.

(c) Category: Window.

(d) Category: Door.

Figure 6: Learning curves for four object categories; our algorithm outperforms a state of the art MIAL algorithm.

Instance Queries	*MIU*	*BU*
Cloud	0.00119	**0.01081**
Street	-0.00105	**0.00483**
Rock	0.0015	**0.0073**
Tree	0.00429	**0.00639**
Window	-0.01243	**0.0097**
Bottle	-0.00456	**0.01009**
Door	-0.02214	**0.00478**
House	**0.01283**	0.001984
Car	-0.00352	**0.00351**
River	-0.00103	**0.00913**

Table 1: The average improvement in AUROC after 50 queries over the baseline MI learner. Numbers are averaged across all trials for each object category. Results corresponding to the algorithm with better performance are indicated in bold font. BU performs better than MIU in all but one class.

sults in significant improvement in classification accuracy in comparison with a state of the art MIAL algorithm.

Future work will investigate the use of different visual features, and evaluate the algorithm on additional object classes. Another direction of further research is to include and reason with some prior domain knowledge about object properties and context, which will facilitate the choice of images presented to humans for feedback. The long-term objective is to implement such algorithms on robots that will be able to interact and collaborate with non-expert humans in complex real-world domains.

Acknowledgments

This work was supported in part by the Office of Naval Research Science of Autonomy award N00014-09-1-0658.

References

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

Cantrell, R.; Scheutz, M.; Schermerhorn, P.; and Wu, X. 2010. Robust Spoken Instruction Understanding for HRI. In *International Conference on Human-Robot Interaction*.

Cover, T. M., and Thomas, J. A. 2006. *Elements of Information Theory, 2nd Edition*. Wiley-Interscience.

Druck, G.; Settles, B.; and McCallum, A. 2009. Active Learning by Labeling Features. In *Empirical Methods in Natural Language Processing*, 81–90.

Escalante, H. J.; Hernndez, C. A.; Gonzalez, J. A.; Lpez-Lpez, A.; Montes, M.; Morales, E. F.; Sucar, L. E.; Villaseor, L.; and Grubinger, M. 2010. The Segmented and Annotated IAPR TC-12 Benchmark. *Computer Vision and Image Understanding* 114(4):419–428.

Fellbaum, C. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11):39–41.

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. In press.

Hawes, N.; Wyatt, J.; Sridharan, M.; Jacobsson, H.; Dearden, R.; Sloman, A.; and Kruijff, G.-J. 2010. Architecture and Representations. In *Cognitive Systems*, volume 8 of *Cognitive Systems Monographs*. Springer Berlin Heidelberg. 51–93.

Marcus, M.; Marcinkeiwicz, M.; and Santorini, B. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. In *Computational Linguistics*.

Maron, O., and Lozano-Perez, T. 1997. A Framework for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems*.

Maron, O., and Ratan, A. L. 1998. Multiple-Instance Learning for Natural Scene Classification. In *International Conference on Machine Learning*.

Nocedal, J., and Wright, S. 1999. *Numerical Optimization*. Springer.

Ratnaparkhi, A. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution. In *PhD dissertation, University of Pennsylvania*.

Settles, B.; Craven, M.; and Ray, S. 2008. Multiple-Instance Active Learning. In *Neural Information Processing Systems*.

Settles, B. 2009. Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Settles, B. 2012. Active Learning. In Ronald J. Brachman and William W. Cohen and Thomas G. Dietterich., ed., *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers.

Siddiquie, B., and Gupta, A. 2010. Beyond Active Noun Tagging: Modeling Contextual Interactions for Multi-Class Active Learning. In *CVPR*.

Swaminathan, R., and Sridharan, M. 2012. Towards Robust Human-Robot Interaction using Multimodal cues. In *Human Agent Robot Teamwork Workshop at the International Conference on Human-Robot Interaction*.

Toutanova, K.; Klein, D.; Manning, C.; and Singer, Y. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *In Proceedings of HLT-NAACL*.

Yang, C., and Lozano-Perez, T. 2000. Image Database Retrieval with Multiple-Instance Learning Techniques. In *International Conference on Data Engineering*.

Zha, Z.-J.; Hua, X.-S.; Mei, T.; Wang, J.; Qi, G.-J.; and Wang, Z. 2008. Joint Multi-Label Multi-Instance Learning for Image Classification. In *CVPR*, 1–8.

Zhang, Q., and Goldman, S. 2001. EM-DD: An Improved Multiple-Instance Learning Technique. In *Neural Information Processing Systems*.

Zhang, D.; He, J.; and Lawrence, R. 2013. MI2LS: Multi-Instance Learning from Multiple Information Sources. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 149–157.