Reasoning about Object Affordances in a Knowledge Base Representation

Yuke Zhu, Alireza Fathi, and Li Fei-Fei

Computer Science Department, Stanford University
{yukez,alireza,feifeili}@cs.stanford.edu

Abstract. Reasoning about objects and their affordances is a fundamental problem for visual intelligence. Most of the previous work casts this problem as a classification task where separate classifiers are trained to label objects, recognize attributes, or assign affordances. In this work, we consider the problem of object affordance reasoning using a knowledge base representation. Diverse information of objects are first harvested from images and other meta-data sources. We then learn a knowledge base (KB) using a Markov Logic Network (MLN). Given the learned KB, we show that a diverse set of visual inference tasks can be done in this unified framework without training separate classifiers, including zero-shot affordance prediction and object recognition given human poses.

1 Introduction

Visual reasoning is one ultimate goal of visual intelligence. Take an apple in Fig. 1 for example. Given a picture of an apple, humans can recognize the object name, its shape, color, texture, infer its taste, and think about how to eat it. Much of our field's effort in visual reasoning is focused on assigning a class label to some part of an image. Indeed casting the reasoning problem as a classification problem is intuitive. Most of the powerful machine learning tools are based on optimizing a classification objective. But this classifier-based paradigm also has limitations. Compared to the rich reasoning that can go through a person's mind upon seeing an apple, a typical object classifier is doing a "shallow" reasoning.

In this paper, we focus on the task of predicting the affordances of objects, and illustrate how a new representation of the visual and semantic information can go beyond this "shallow" reasoning and allow for more flexible and deeper visual reasoning. Gibson in his seminal paper [16] in 1979 refers to affordance as "properties of an object [...] that determine what actions a human can perform on them." Inspired by this, and a number of recent studies in computer vision [17,21,18,37], we define the full description of affordance as a combination of three things: (1) an affordance label (e.g. edible), (2) a human pose representation of the action (e.g. in skeleton form) and (3) a relative position of the object with respect to the human pose (e.g. next to).

A Naïve Approach. One way to make a rich prediction of affordance is to train a battery of different classifiers, each focusing on one aspect (color, shape,

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 408–424, 2014.

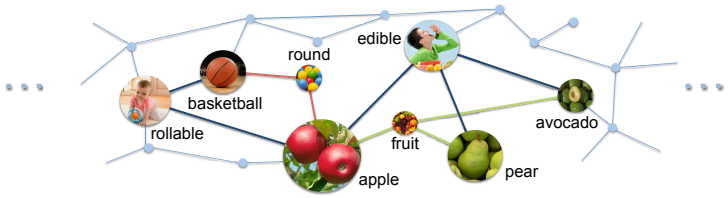


Fig. 1. An example knowledge structure for visual reasoning. Relevant nodes are interconnected in the knowledge graph. Different types of edges (indicated by color) depict a diverse range of relations between nodes, which relate different concepts, such as objects, their attributes and affordances, to each other.

texture, etc.) of the object. However, assuming we can do this for an exhaustive list of attributes of an apple, the reverse question remains — namely, inferring the type of fruit given an image of a person eating a piece of fruit or an image of a red, round piece of fruit.

Knowledge-Based Approach. Another way to consider the problem of visual reasoning is through a knowledge structure, such as the one illustrated in Fig. 1. Apple, in a knowledge graph, is a node (or entity) connected to other nodes, some depicting its visual attributes such as shape, color, texture, and other nodes depicting its affordance, such as edible. Each node connecting apple is further connected to other relevant nodes. In Fig. 1, the edible node is connected with pear, and the round node is connected to basketball, etc.

This representation is well known in the database and NLP communities, often called *knowledge base* (KB) or *knowledge graph*. Compared to classifiers that tackle one specific task, using a knowledge-based representation can enable querying a much larger array of questions. In one unified system, once the building and training the KB is complete, we are able to perform tasks such as zero-shot inference of object affordances, estimation of action pose given a visual object, prediction of an object given a likely action, etc. When using the aforementioned naïve approach, we would have trained separate classifiers for each of these tasks, each requiring a different set of training data and labels.

This paper presents a principled way of building a knowledge base (KB) of objects, their attributes, and affordances by extracting information from images as well as online textual sources such as Amazon and eBay. We use a Markov Logic Network (MLN) model [28] to represent the KB. We emphasize that once the KB of objects and their properties are trained, we can perform a number of different inference tasks in a unified framework without any further training. We demonstrate the effectiveness of this representation by testing on a number of sub-tasks related to zero-shot object affordance inference as well as object prediction given human poses. Our system outperforms classifiers trained for each individual sub-task.

2 Previous Work

Object Affordances. While the majority of visual recognition work focus on learning visual appearance based classifiers of objects [12,11,23], there is a

growing interest in recognizing object and scene affordances (some call "functionalities") [21,17,15,37,20,22,19]. Winston et al. [34] learn physical descriptions of objects from their functional definitions. Gupta and Davis [18] and Kjellström et al. [21] use the functionality to detect objects. Grabner et al. [17] and Jiang et al. [20] represent affordances by hallucinating humans as the hidden context. Yao et al. [37] represent the functionality of an object based on the majority's human poses during interactions with it. None of the work, however, can predict affordance in novel objects. Furthermore, most of these work predict affordance as a single label whereas we can simultaneously predict affordance label, human pose and relative object location in a unified framework.

Zero-Shot Learning of Objects and Attributes. The classic approach to recognizing unseen objects is based on the visual similarity of the novel object with previously seen examples (e.g. [14,2]). More recently, Lampert et al. [25] introduced a method that recognizes unseen objects by transferring attributes from previously observed classes. Parikh and Grauman [27] extend this work by replacing binary attributes with relative attributes. Rohrbach et al. [29] compare three methods for knowledge transfer: object similarity, attributes, and object hierarchy. Furthermore, they mine attributes from the web to improve the performance of their method. In contrast to these methods, (a) we can predict affordances of unseen objects and infer much richer information beyond visual similarities, and (b) we use a knowledge based approach for reasoning and answering various types of queries, both through images and text.

Knowledge Base Representation. There is a growing trend towards building large-scale knowledge bases with statistical learning methods. NELL [5] learns probabilistic horn clauses by extracting and analyzing information from web text. NEIL [6] is a framework to automatically extract common sense relationships from web images. The *Jeopardy!*-winning DeepQA project [13] proposes a probabilistic evidence-based question-answering architecture involving more than 100 different techniques. Similar to this work, StatSnowball [38] and Elementary [26] use Markov Logic Networks [28] as the underlying knowledge representation and perform statistical inference for knowledge base construction. Tran and Davis [33] use Markov Logic Network to model events that contain complex interactions of people and vehicles. In contrast to these models, our knowledge base incorporates a wide range of heterogeneous information, allowing us to answer a diverse set of visual and textual queries.

3 Knowledge Base Construction and Representation

We first present our method for constructing a KB that relates objects, their attributes, and their affordances comprised of the three aforementioned components (affordance labels, human poses and human-object relative locations). To illustrate our idea, we use 40 objects and their properties for constructing the KB. However, our method is scalable to an arbitrary number of objects.

3.1 Overview of the Knowledge Base

A knowledge base (KB) refers to a repository of entities and rules that can be used for problem solving. One can also think of the KB as a graph (similar to Fig. 1), where the nodes denote the entities and the edges, denoting the general rules, characterize their relations.

Entities. The entities in our KB consist of object attributes and affordances. We use three types of attributes to describe an object:

1. **Visual attributes** – correspond to knowledge acquired from visual perception. Inspired by recent work on attribute learning [9,25,27], we define a set of visual attributes as a mid-level description of visual appearance.
2. **Physical attributes** – constitute a form of knowledge from the physical world. Each physical attribute is a measurable quantity that describes one aspect of the object. We select two relevant properties, weight and size, to describe the objects.
3. **Categorical attributes** – reflect a semantic understanding (generalization) of the object. Object categories form a hierarchy consisting of several levels of abstraction [8]. Knowledge of categorical attributes (e.g. a *dog* is an *animal*) often facilitates the ability of affordance reasoning.

These attributes serve as an intermediate representation of objects. This representation allows us to transfer knowledge across objects, and thus to predict the affordances of an object even if it has never seen before, which are represented by three types of entities:

1. **Affordance labels** – a verb or a verb phrase (e.g. *ride* and *sit on*).
2. **Human poses** – an articulated skeleton of human poses.
3. **Huamn-object relative locations** – the spatial relations between the human and the object during a human-object interaction.

General Rules. The general rules describe the relations between the entities. One can think of them as the edges in the knowledge graph. We model three types of relations between these entities:

1. **Attribute-Attribute Relations.** Strong correlations exist between attributes. We model these correlations with *attribute-attribute* relations. Positive weights indicate a positive correlation between two attributes; conversely, negative weights indicate that these attributes are not likely to co-occur.
2. **Attribute-Affordance Relations**. We observe that the affordances of an object are largely dependent upon its attributes (e.g. laptops and umbrellas are *lift-able* because they are not *heavy*). We model these dependency relationships by a set of *attribute-affordance* relations.

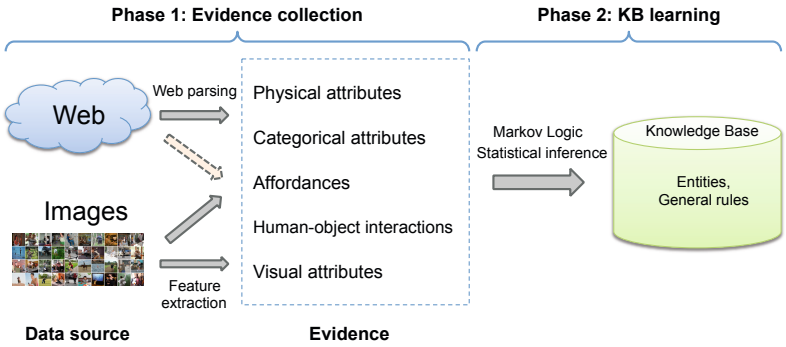


Fig. 2. A system overview of knowledge base learning. This process consists of two phases (Section 3.2). First we collect the evidence from diverse data sources, including images and online text. Then we learn the KB using Markov Logic Network.

3. **Human-Object-Interaction (HOI) Relations.** Humans are likely to interact with different objects in different ways. Furthermore, an object's attributes affect the way that a human interacts with it (e.g. the weight of an object changes the way it is grasped). Therefore, human poses and human-object spatial relations are jointly determined by the object attributes and the affordance. We define four sets of HOI relations to model the correlations (*attribute-pose*, *affordance-pose*, *attribute-location* and *affordance-location*).

3.2 Learning the Knowledge Base

Now that we have defined the entities and rules of the KB, we are ready to learn it from source data. There are two phases in learning the KB. First we collect evidence from diverse sources containing images and online textual sources. Then we employ Markov Logic Network (MLN) [28] for knowledge representation. Fig. 2 is a system overview of the key steps in the learning process. We now elaborate each of these steps below.

Phase 1: Collecting Evidence for KB Construction. A KB is populated by evidence, a set of facts and assertions about the entities. As Fig. 1 and Fig. 2 illustrate, we would like our KB to incorporate a wide range of heterogeneous information, including object attributes, affordances, human poses, etc.

Data source — We choose 40 objects offered by Stanford 40 Actions dataset [36] to seed the KB. For each object, we sample 100 images from the ImageNet dataset [7]. We select 14 affordances from human actions in Stanford 40 Actions. Fig. 3 shows 10 out of the 40 objects and the 14 different affordance labels. On average, each object has 4.25 out of the 14 affordances. Note that, the first four affordances are low-level physical interactions, which are a major interest in the robotics community; while, the rest are daily actions that often involve more complex human-object interaction and demand a higher-level understanding.

Evidence — Given the 40 objects, we are now ready to collect a set of evidence for the KB from the images as well as a number of online sources, such

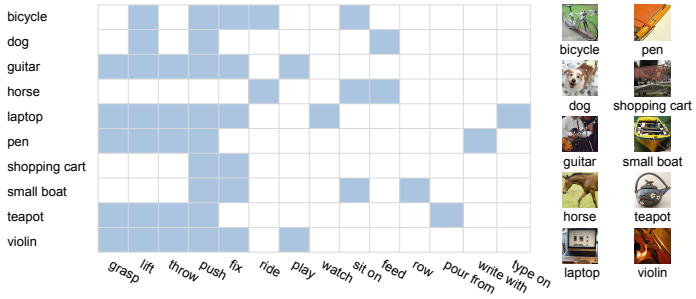


Fig. 3. Object images and affordance labels. We illustrate 10 objects in our KB and their affordance labels. The x-axis lists the 14 affordances and the y-axis provides the names of the 10 objects with sample images on the right. The presence of an affordance is indicated by blue color.

as Freebase [3], WordNet [10] and online shopping sites. For constructing a good KB, we would like the evidence to be diverse, accurate and consistent.

1. **Visual attributes.** Following [9], we choose 33 pre-trained visual attribute classifiers[1] to describe the shape, material and parts of the objects.
2. **Physical attributes.** We extract the real-world weights and real-world sizes of the objects from the *animal synopsis* fields on Freebase [3] and *product details* data from Amazon[2] and eBay[3]. To accommodate the noise in web data, we take the medians of the top K retrieved results as the true values. We quantize the weights into four bins (<1kg, 1–10kg, 10–100kg and >100kg) and the sizes into three bins (<10in, 10–100in and >100in). Fig. 4 shows a list of objects ranked by their weights and the four bins for quantization.

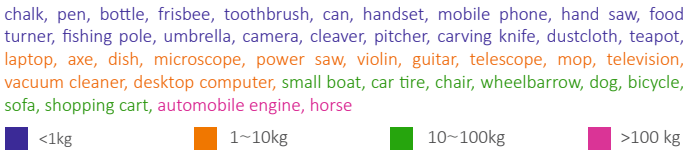


Fig. 4. (Best viewed in color) Objects ranked by their physical weights. The weights are automatically collected from web sources. The quantization bins are indicated by the font colors. These estimates roughly reflect real-world weights of objects. Some objects (e.g. toothbrush and dustcloth) get larger estimates than expected since they are usually sold in batch on the shopping sites we use as data sources.

[1] **Visual attributes:** boxy 2D, boxy 3D, clear, cloth, feather, furn. arm, furn. back, furn. leg, furn. seat, furry, glass, handlebars, head, horiz. cyl., label, leather, metal, pedal, plastic, pot, rein, round, saddle, screen, shiny, skin, tail, text, vegetation, vert. cyl., wheel, wood, wool

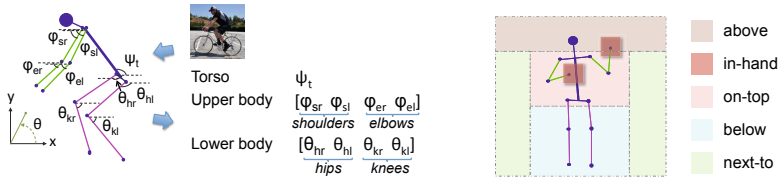
[2] `http://www.amazon.com/`

[3] `http://www.ebay.com/`

3. **Categorical attributes.** Membership to more general classes can be informative for object reasoning [24,8]. We refer to this as categorical attributes. We obtain these attributes by extracting the hypernym hierarchy from lexical ontologies such as WordNet [10]. The hypernyms of an object can be regarded as a generalization of this object (e.g. hypernyms of *dog* are *mammal*, *animal*, etc.). To improve computational efficiency, we merge hypernyms that cover the same set of objects and remove those containing only one object. Finally, we use 22 hypernyms[4] as categorical attributes.
4. **Affordance labels.** As Fig. 3 illustrates, multiple affordance labels are assigned to each object in the KB. For this paper, we provide a manual labeling of the affordances to the 40 objects used for training. But one alternative approach would be to obtain a few canonical affordances is to extract the most frequent verbs associated with a noun phrase in large corpus like Google N-gram (the dashed arrow in Fig. 2).
5. **Human poses.** Human poses can be extracted from human action images in Stanford 40 Actions. Many approaches [35,1,37] have been proposed for this task, yet the state-of-the-art methods fail to perform robustly on images with large variations. To ensure the robustness of our KB in the training phase, we annotate the human poses of the images manually. We compute a pose descriptor based on the tilt angles of body parts (see Fig. 5(a)). The body part descriptors are discretized by k-means. The number of cluster centroids is determined by the Elbow Method. In practice, we choose 3 clusters for torsos, 8 for lower bodies and 8 for upper bodies.
6. **Human-object relative locations.** We extract human-object spatial relations based on the relative locations and sizes of their bounding boxes from human action images. The spatial relations are quantized into five bins: *above*, *on-top*, *below*, *next-to* and *in-hand* (see Fig. 5(b)).

Phase 2: Learning the KB Using Markov Logic. Given the collected evidence, we build the KB by learning the relations, i.e. the weights of the general rules. We employ a Markov Logic Network (MLN) [28] for knowledge representation. Fig. 6 summarizes the schema and general rules with some examples. The idea of MLN is to unify Markov Random Fields (MRF) and first-order logic. Markov Logic is a widely used language in statistical relational learning, which specifies an MRF by a weighted first-order logic knowledge base. Learning and inference in MLN resemble the standard algorithms for MRF, where a ground MRF is first instantiated by the weighted logic formulae. The formulae representing the entities and general rules define the structure of the KB. MLN can be considered as a log-linear model with one node per ground atom and one feature

[4] **Categorical attributes:** animal, instrumentality, implement, device, container, tool, equipment, vehicle, machine, wheeled vehicle, vessel, electronic equipment, edge tool, handcart, seat, musical instrument, cooking utensil, computer, scientific instrument, knife, telephone, writing implement



(a) pose descriptor (b) relative locations

Fig. 5. Pose descriptors and human-object relative locations. (a) Human pose is represented by the tilt angles of body parts. The upper bodies are described by the angles of left and right shoulders and elbows, and the lower bodies by the angles of hips and knees. (b) Relative object locations are represented as quantized bins based on the centers and sizes of their bounding boxes. We use a total of five spatial bins to describe the human-object spatial relations.

Schema	General Rules	Examples
hasAffordance(*object*, *affordance*)	Attribute-attribute relations	isA(x, Vehicle) ⇒ isA(x, Animal)
isA(*object*, *category*)		
hasVisualAttribute(*object*, *attribute*)	Attribute-affordance relations	hasVisualAttribute(x, Furry) ⇒ hasAffordance(x, Feed)
hasWeight(*object*, *weight*)		hasWeight(x, W4) ⇒ hasAffordance(x, SitOn)
hasSize(*object*, *size*)		
locate(*object*, *location*)	Human-object-interaction relations	hasAffordance(x, Ride) ∧ locate(x, Below)
torso(*object*, *torso_id*)		isA(x, Animal) ∧ locate(x, Below)
upperBody(*object*, *ubody_id*)		hasAffordance(x, Push) ∧ torso(x, T1)
lowerBody(*object*, *lbody_id*)		isA(x, Vehicle) ∧ upperBody(x, U3)

Fig. 6. Knowledge base schema and general rules. The arguments in the schema specify the types of the variables. $W2$, $T1$ and $U3$ correspond to the quantized object weight (1–10kg), the first cluster for the torso and the third cluster for the upper body.

per ground formula. The joint distribution over possible worlds x is given by

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_{i=1}^{n} w_i f_i(x_{\{i\}})\right) \tag{1}$$

where Z is the partition function, F is the set of first-order formulae in MLN and n is the number of formulae in F, $x_{\{i\}}$ is a state of ground atoms appearing in the formula F_i and the feature function $f_i(x_{\{i\}}) = 1$ if $F_i(x_{\{i\}})$ is true and 0 otherwise. The weights w indicate the likelihood of the formulae being true. We learn the optimal weights w^* by maximizing the pseudo-likelihood given the evidence collected in Section 3.2 using the L-BFGS algorithm [28].

3.3 Visualizing the Knowledge Base

Fig. 7 visualizes a part of the constructed knowledge base. In this graph, each node (entity) corresponds to an atomic formula in MLN, and each edge (general rule) corresponds to a first-order logic formula that composes two atomic formulae with logic connectives and quantifiers. The weights of the edges are learned in Markov Logic (Section 3.2), where positive weights indicate that two entities are likely to co-occur (e.g. *furry* and *feed*), and negative weights indicate the entities are negatively correlated (e.g. *fix* and *animal*).

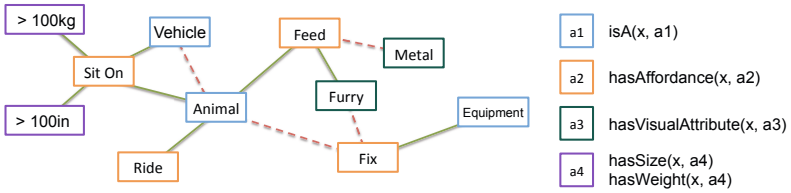


Fig. 7. Graphical illustration of the constructed KB. The nodes denote the entities (atomic formulae in MLN) illustrated on the right. The edges denote the *attribute-attribute* and *attribute-affordance* relations. The green solid edges indicate positive weights and the red dashed edges indicate negative weights.

0.8232	hasVisualAttribute(x, Saddle) ⇒ hasAffordance(x, SitOn)
0.7467	hasVisualAttribute(x, Pedal) ⇒ hasAffordance(x, Lift)
0.7155	hasVisualAttribute(x, Screen) ⇒ hasAffordance(x, Fix)
0.7012	hasVisualAttribute(x, Head) ⇒ hasAffordance(x, Feed)
0.6540	hasVisualAttribute(x, Furry) ⇒ hasAffordance(x, Feed)

(a) Top positive attributes (Visual)

-1.0682	hasVisualAttribute(x, Metal) ⇒ hasAffordance(x, Feed)
-1.0433	hasVisualAttribute(x, Shiny) ⇒ hasAffordance(x, Feed)
-1.0115	hasVisualAttribute(x, Boxy_3D) ⇒ hasAffordance(x, Feed)
-0.8317	hasVisualAttribute(x, Wheel) ⇒ hasAffordance(x, Feed)
-0.7987	hasVisualAttribute(x, Text) ⇒ hasAffordance(x, Feed)

(b) Top negative attributes (Visual)

5.4734	isA(x, Animal) ⇒ hasAffordance(x, Feed)
3.3196	isA(x, Vehicle) ⇒ hasAffordance(x, Ride)
3.2436	isA(x, Vehicle) ⇒ hasAffordance(x, Row)
2.7976	isA(x, Container) ⇒ hasAffordance(x, PourFrom)
2.6208	isA(x, Animal) ⇒ hasAffordance(x, SitOn)

(c) Top positive rules

-3.8636	isA(x, Animal) ⇒ hasAffordance(x, Fix)
-2.2209	isA(x, Seat) ⇒ hasAffordance(x, Push)
-1.8066	isA(x, Vehicle) ⇒ hasAffordance(x, Lift)
-1.7254	isA(x, Instrumentality) ⇒ hasAffordance(x, Feed)
-1.3258	isA(x, Instrumentality) ⇒ hasAffordance(x, Fix)

(d) Top negative rules

Fig. 8. Top weighted attribute-affordance relations. The relations between categorical attributes and affordances have the largest weights, indicating their importance in determining object affordances. For comparison, we also provide the top weighted relations between visual attributes and affordances. The weighted rules can be well interpreted. For instance, the first rule in Fig. 8(b) denotes that "objects that look metal are less likely to be feed-able".

To ensure the quality of the KB, we further examine the weights of general rules learned by MLN statistical inference. Large positive/negative weights indicate a high confidence of the rule being true/false [28]. Fig. 8 lists the top positive and negative weighted attribute-affordance relations. In contrast to visual attributes, categorical attributes serve as a more discriminative semantic-level abstraction, and therefore have larger weights.

4 Affordance Reasoning with KB

Now that we have learned a KB containing rich information about objects, their attributes and affordances, we show in this section a number of experiments to illustrate the effectiveness of this knowledge representation. We emphasize on the word *reasoning*. One of the most important advantages of using a KB representation is to allow for different types of visual and textual queries in a unified framework, as opposed to training separate classifiers for each task. Section 4.1 and Section 4.2 show experimental results for a number of visual tasks. Section 4.3 further explores some important properties of the KB.

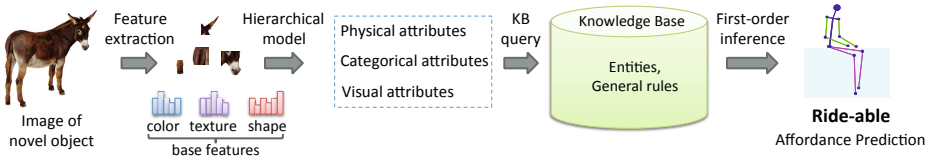


Fig. 9. The inference procedure of zero-shot affordance prediction. Given an image of a novel object, our model estimates the object attributes via a hierarchical model. These attributes serve as evidence for KB queries. We then employ first-order probabilistic inference to predict the affordances and to estimate human poses and human-object relative locations.

4.1 Zero-Shot Affordance Prediction

Given an unseen object, it is often useful to predict its affordances for both humans and robots. We remind readers that by *affordance*, we mean a combination of three pieces of information: an affordance label, the human pose and human-object relative location. We first briefly discuss the inference procedure, the testing data, and then show a number of experimental results.

Inference. Using the constructed KB, we propose a hierarchical model to perform affordance prediction. Given an image of a novel object, our model employs the visual information as cues to object attributes. The model first estimates visual attributes of the object, and infers its physical and categorical attributes. These attributes are taken as evidence to query the affordances and most likely human poses and object relative locations. We use lifted belief propagation for inference [31]. Fig. 9 illustrates an overview of the inference procedure.

Given an image I, we first extract the base features suggested in [9] and predict visual attributes. We then train a L1-regularized logistic regression classifier for each categorical attribute with both base features and visual attributes. Once we obtain the scores of visual and categorical attributes, we map the scores into a binary vector, where the nonzero entries indicate the presence of these attributes.

We predict the physical attributes by learning a ranking function. Based on the physical attributes of the training objects (see Fig. 4), we construct a set $\mathcal{P}_k$ of pairwise preferences where $(i, j) \in \mathcal{P}_k$ indicates i has a larger value than j of the k-th physical attribute. Our goal is to learn a ranking function $R_k(I) = w_k^T \phi(I)$ that attempts to satisfy $R_k(I_i) > R_k(I_j)\ \forall (i, j) \in \mathcal{P}_k$, where w_k is a model parameter and $\phi(I)$ is the base features. We train the model parameters using the ranking SVM formulation in [27]. Given a novel object, we estimate its physical attributes by comparing its ranking scores to the average scores of training objects.

Testing Data. Based on the 40 objects in the KB, we select a different set of 22 semantically similar objects[5] (close synsets in WordNet hierarchy) for testing.

[5] **Testing objects:** banjo, bench, bowl, broom, camel, cat, coffee cup, donkey, flagon, hammer, hand truck, kayak, monitor, motorcycle, pencil, rhinoceros, serving cart, sickle, spoon, stool, typewriter, walkie-talkie.

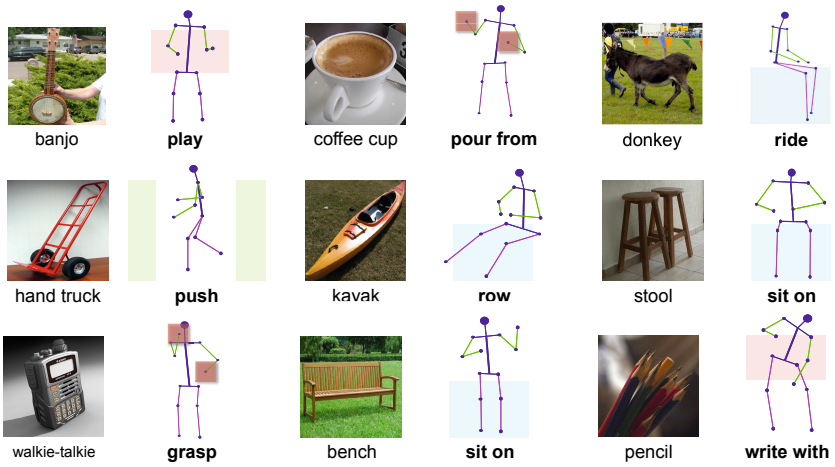


Fig. 10. Results of affordance prediction. We visualize the predicted affordances for a few testing objects. The relative locations are indicated by the color boxes defined in Fig. 5. The first seven examples are correct predictions; while the last two examples fail to match the ground-truth poses/locations.

For each object, we randomly sample 50 images from ImageNet [7]. These images of novel objects are taken as the inputs for affordance prediction.

Result 1: Predicting Affordance Labels. Some results are shown in Fig. 10. Our model can correctly predict the affordances of a novel object given object images from various viewpoints. Besides assigning affordance labels, the model simultaneously estimates the human poses and object relative locations.

For quantitative evaluation, we use the mean *area under the ROC curve* (mAUC) over all the affordances to evaluate the performance of our model (Table 1). We compare our method with two attribute-based classifiers based on previous work [9,25]. From the hierarchical model, we extract the base features and estimate the object attributes. Following [9] and [25], we train linear classifiers with L1-regularized logistic regression and SVMs with a multi-channel χ^2-kernel on four types of features: base features (BF), visual attributes (VA) and categorical & physical attributes (CP) and combined attributes (VA+CP).

Table 1. Performance of Zero-shot Affordance Prediction (measured in mAUC)

Method	**L1-LR** [9]	χ^2**-SVM** [25]	**Ours**
base features (BF)	0.7858	-	-
visual attributes (VA)	0.7525	0.7533	0.7432
categorical & physical (CP)	0.7919	0.7924	0.8234
combined (VA+CP)	0.8006	0.7985	**0.8409**

Our results in Table 1 indicate a combination of features achieve the best performance for the classifiers. In comparison, the knowledge-based model achieves

the best performance with a 4% improvement over the best classification-based method. We attribute the better performance of the knowledge-based model to the complex general rules. Such relations can be readily represented in Markov Logic; however, classifiers fail to take the correlations into account.

Result 2: Estimating Human Poses. We now evaluate how our model predicts the poses of its canonical affordance. Each pose can be represented as a triple index $\mathcal{T}$ of the cluster centroids of the torso, lower body and upper body. We compute a Hamming distance between the index and its nearest neighbor in the ground-truth poses:

$$hamming(\mathcal{T}) = \min_{\mathcal{T}' \in P_o \cup \hat{P_o}} \sum_{i=1..3} \mathbb{1}(\mathcal{T}_i = \mathcal{T}'_i) \tag{2}$$

where P_o and its horizontal mirroring $\hat{P}_o$ are the set of ground-truth poses of the canonical affordance of the object, $\mathcal{T}_i$ is an index of the cluster centroid, and $\mathbb{1}(\cdot)$ is an indicator function. This distance metric ranges from 0 to 3, and a smaller value implies a better estimation of the poses.

Table 2. Performance of Estimating Human Poses (in Hamming distance)

Method	nearest neighbor	attributes	affordances	attributes+affordances
Distance	0.928	1.027	0.630	**0.527**

We compare our method with a nearest neighbor baseline, where we assign the canonical affordance and a corresponding human pose of its nearest neighbor to a testing object. The nearest neighbors are defined upon the Euclidean distance between the VA+CP attributes. We report the mean Hamming distance over all the testing samples in Table 2. To see how attributes and affordances affect the performance, we compare it with two methods, where we provide only the attributes and the affordances as evidence respectively. The best performance is achieved by combining affordances and attributes together. However, using affordances alone significantly outperforms its attribute counterpart. This may be due to the limited number of objects in the KB that have a certain affordance; thus in many cases, it is sufficient to predict the poses given the affordance.

4.2 Prediction from Human-Object Interactions

A reverse direction towards affordance prediction is to recognize the action and hypothesize the object in human-object interactions. When actions are seen at a distance and objects appear small, it is hard to observe object's visual attributes. In such cases, human poses and human-object spatial relations provide complementary information. We demonstrate the effectiveness of our KB in predicting the actions and the objects from human-object interactions.

Inference. From human action images, we extract the quantized human poses as evidence and query the affordance labels as well as object attributes. The affordance label with the highest likelihood is taken as the predicted action. We perform Maximum *a posteriori* (MAP) inference on MLN to estimate the most

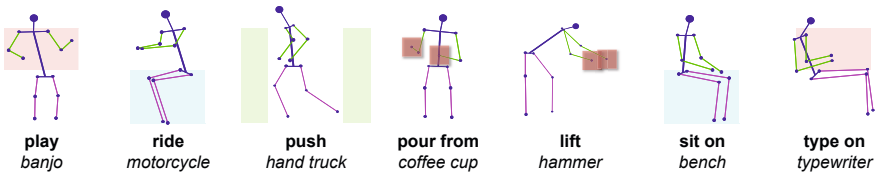


Fig. 11. Prediction results from human-object interactions. We provide some examples of correct predictions of the actions and the objects. The last two examples illustrate two similar poses. Both are predicted as *sit on bench* using only the poses. The relative locations in the full model disambiguate those two poses.

likely 0/1 state of each object attribute. The predicted attributes can be used to retrieve the nearest neighbor among all the testing objects in Euclidean distance. We further evaluate how human-object spatial relations affect the performance. The relative locations between humans and objects are extracted from human action images as described in Fig. 5. We add the quantized locations in the evidence and perform the same queries.

Testing Data. We collect five human action images for each of the 22 testing objects from Stanford 40 Actions [36] and Google Image Search. We focus on one canonical affordance (e.g. *riding* for *motorcycle*) for each object.

Results. Fig. 11 provides some prediction results. Our model utilizes the information of the poses and relative locations to predict the actions and the objects.

We use prediction accuracy to quantitatively measure whether the model is able to correctly predict the action and the object in Table 3. One can see that human poses provide useful information about the actions. However, poses alone are sometimes insufficient to characterize an action. Human-object spatial relations disambiguate similar poses and therefore boost the performance. Besides, our model works better in predicting the affordance labels than the objects. In cases where humans interact with objects in similar ways, it is hard to tell apart objects but easier to identify the actions.

Table 3. Predicting Actions and Objects from Human-Object Interactions

Method	Action	Object
human poses	50.4%	46.2%
poses + locations	81.2%	64.5%

4.3 Why KB - Empirical Results

Partial Observation. Humans are proficient in inferring information given a few clues. For instance, people can easily identify a gray, heavy animal with a long trunk as an elephant. The ability of reasoning from partial observation is derived from the knowledge that connects the dots of various concepts together. In this section, we demonstrate the robustness of our model against partial observation.

To evaluate the robustness of our model, we test the performance of our model in affordance prediction given a randomly selected portion of evidence. For comparison, we evaluate the performance of the classification-based method. During testing, a portion of attribute feature dimensions are randomly selected and zeroed out. Fig. 12 depicts how the performance (mAUC) varies as a larger portion of attributes become unobserved.

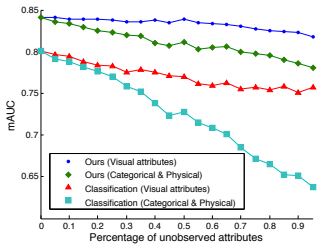


Fig. 12. Performance variations against partial observation. The x-axis denotes the percentage of unobserved evidence. The y-axis denotes the performance (mAUC). The top two curves correspond to our method. The bottom two are the classification-based method. In comparison, the knowledge base representation is more robust against partial observation.

We observe that performance drops significantly as information of categorical and physical attributes get removed; while, both models are relatively robust against the removal of visual attributes. In general, the performance of classification-based model drops more rapidly than the knowledge-based approach. This result provides evidence for KB's ability to utilize its rich structure for inference against partial observations, while there is no such mechanism built in traditional classifiers.

Diverse Question Answering. Compared to the classifiers, a KB representation can enable querying and inferencing of a much larger array of questions. Given a set of weighted MLN formulae, a user may write arbitrary queries in

Question	Evidence	Query	Top Answers
What do animals look like?	isA(N1, **Animal**)	hasVisualAttribute(N1, x)	hasVisualAttribute(N1, **Leather**) hasVisualAttribute(N1, **Head**) hasVisualAttribute(N1, **Tail**) hasVisualAttribute(N1, **Furry**)
I saw something shiny and metallic. What is it?	hasVisualAttribute(N1, **Shiny**) hasVisualAttribute(N1, **Metal**)	isA(N1, x)	isA(N1, **Instrumentality**) isA(N1, **Device**) isA(N1, **Container**) isA(N1, **Computer**)
Here is a vehicle and it's quite heavy. What can I do with it?	isA(N1, **Vehicle**) hasWeight(N1, W4) **(> 100 kg)**	hasAffordance(N1, x)	hasAffordance(N1, **Ride**) hasAffordance(N1, **Row**) hasAffordance(N1, **SitOn**) hasAffordance(N1, **Fix**)
Tell me how heavy and large a wooden musical instrument is.	isA(N1, **Musical_instrument**) hasVisualAttribute(N1, **Wood**)	hasWeight(N1, x) hasSize(N1, x)	hasSize(N1, D2) **(10-100 in)** hasWeight(N1, W2) **(1-10 kg)**

Fig. 13. Examples of question answering. We convert each question into the form of evidence and queries, where $N1$ is used for grounding. Predicates with the highest probabilities computed from MLN inference are presented in the last column as answers to the queries.

terms of the entities and rules. To answer these queries, MLN infers the probability or the most likely state of each query from the evidence. In Fig. 13, we provide examples to show the power of the KB in diverse question answering. Note that in a unified framework, we are able to query with both textual (e.g. *isA*) and visual (e.g. *hasVisualAttribute*) questions. Furthermore, the answers returned by the KB can also be textual (e.g. *hasSize*) or visual (e.g. *hasVisualAttribute*).

5 Conclusion

In this paper, we presented a knowledge-based (KB) representation to reason about objects, and their affordances in human-object interactions, motivated by a need to conduct deeper and more diverse reasoning of the heterogeneous data in the form of images and text. Our preliminary results show that a KB representation is a powerful tool to organize the rich information of the visual world, and to allow us to query different types of questions related to objects and their affordances, compared to a number of traditional classification schemes. A natural future direction is to extend the KB into a much larger scale for richer inferences. In this work, we choose to express our data structure and inference in a Markov Logic Network (MLN). A number of recent advances in database and machine learning [30,32,4] also point ways to different inference algorithms.

Acknowledgement. This work is partially supported by an NSF CAREER grant (IIS-0845230), an ONR MURI grant, the DARPA VIRAT program and the DARPA Mind's Eye program. We would like to thank Kevin Tang, Andrej Karpathy, Justin Johnson and anonymous reviewers for useful comments.

References

1. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: CVPR (2009)
2. Bart, E., Ullman, S.: Single-example learning of novel classes using representation by similarity. In: BMVC (2005)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A collaboratively created graph database for structuring human knowledge. In: ACM SIGMOD International Conference on Management of Data (2008)
4. Bordes, A., Weston, J., Collobert, R., Bengio, Y.: Learning structured embeddings of knowledge bases. In: AAAI Conference on Artificial Intelligence (2011)
5. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI Conference on Artificial Intelligence (2010)
6. Chen, X., Shrivastava, A., Gupta, A.: Neil: Extracting visual knowledge from web data. In: IEEE International Conference on Computer Vision (2013)
7. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE International Conference on Computer Vision (2009)

8. Deng, J., Krause, J., Berg, A.C., Fei-Fei, L.: Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: Computer Vision and Pattern Recognition (2012)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: Computer Vision and Pattern Recognition (2009)
10. Fellbaum, C.: Wordnet: An electronic lexical database. Bradford Books (1998)
11. Felzenszwalb, P., McAllester, D., Ramaman, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
12. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: ICCV (2005)
13. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building watson: An overview of the deepqa project. AI Magazine (2010)
14. Fink, M.: Object classification from a single example utilizing class relevance pseudo-metrics. In: NIPS (2004)
15. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: human actions as a cue for single view geometry. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 732–745. Springer, Heidelberg (2012)
16. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
17. Grabner, H., Gall, J., Gool, L.V.: What makes a chair a chair? In: CVPR (2011)
18. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: using spatial and functional compatibility for recognition. PAMI (2009)
19. Gupta, A., Satkin, S., Efros, A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR (2011)
20. Jiang, Y., Koppula, H.S., Saxena, A.: Hallucinated humans as the hidden context for labeling 3d scenes. In: CVPR (2013)
21. Kjellstrom, H., Romero, J., Kragic, D.: Visual object action recognition: inferring object affordances from human demonstration. In: CVIU (2010)
22. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. In: Robotics: Science and Systems (RSS) (2013)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
24. Kuettel, D., Guillaumin, M., Ferrari, V.: Segmentation propagation in imagenet. In: European Conference on Computer Vision (2012)
25. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
26. Niu, F., Zhang, C., Ré, C., Shavlik, J.: Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. In: International Journal on Semantic Web and Information Systems - Special Issue on Web-Scale Knowledge Extraction (2012)
27. Parikh, D., Grauman, K.: Relative attributes. In: International Conference on Computer Vision (2011)
28. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning 62(1-2), 107–136 (2006)
29. Rohrbach, M., Stark, M., Schiele, B.: Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In: CVPR (2011)
30. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008)

31. Singla, P., Domingos, P.: Lifted first-order belief propagation. In: AAAI Conference on Artificial Intelligence (2008)
32. Socher, R., Chen, D., Manning, C.D., Ng, A.Y.: Reasoning with neural tensor networks for knowledge base completion. In: Conference on Neural Information Processing Systems (2013)
33. Tran, S.D., Davis, L.S.: Event modeling and recognition using markov logic networks. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 610–623. Springer, Heidelberg (2008)
34. Winston, P.H., Binford, T.O., Katz, B., Lowry, M.: Learning physical descriptions from functional definitions, examples, and precedents. In: AI Memos (1982)
35. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures of parts. In: CVPR (2011)
36. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: IEEE International Conference on Computer Vision (2011)
37. Yao, B., Ma, J., Fei-Fei, L.: Discovering object functionality. In: ICCV (2013)
38. Zhu, J., Nie, Z., Liu, X., Zhang, B., Wen, J.-R.: Statsnowball: a statistical approach to extracting entity relationships. In: International World Wide Web Conference (2009)