Mining Semantic Affordances of Visual Object Categories

Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng

Computer Science & Engineering, University of Michigan, Ann Arbor
{ywchao, wangzhan, mihalcea, jiadeng}@umich.edu

Abstract

Affordances are fundamental attributes of objects. Affordances reveal the functionalities of objects and the possible actions that can be performed on them. Understanding affordances is crucial for recognizing human activities in visual data and for robots to interact with the world. In this paper we introduce the new problem of mining the knowledge of semantic affordance: given an object, determining whether an action can be performed on it. This is equivalent to connecting verb nodes and noun nodes in WordNet, or filling an affordance matrix encoding the plausibility of each action-object pair. We introduce a new benchmark with crowdsourced ground truth affordances on 20 PASCAL VOC object classes and 957 action classes. We explore a number of approaches including text mining, visual mining, and collaborative filtering. Our analyses yield a number of significant insights that reveal the most effective ways of collecting knowledge of semantic affordances.

1. Introduction

Affordances are fundamental attributes of objects. Affordances reveal the functionalities of objects and the possible actions that can be performed on them. An object is a chair because it affords the possibility to be sit on. An object is a food item because it is edible. Understanding affordances is crucial for recognizing human activities in images and videos because the functionality of objects informs us about possibile activities and roles—a person wearing a stethoscope is likely to be a doctor; a person looking at a clock is likely to be checking the time. In addition to helping computers better undersand human activities, the knowledge of affordances is also essential for a robot to interact with the environment and achieve its goals.

The key question is: given an object, can an action be performed on it? Can a dog be hugged? What about an ant? Can a TV be turned on? What about a bottle? While these questions might seem obvious to a human, to the best of our knowledge, there is no automated system that can readily answer this question and there is no knowledge base that provides comprehensive knowledge of object affordances.

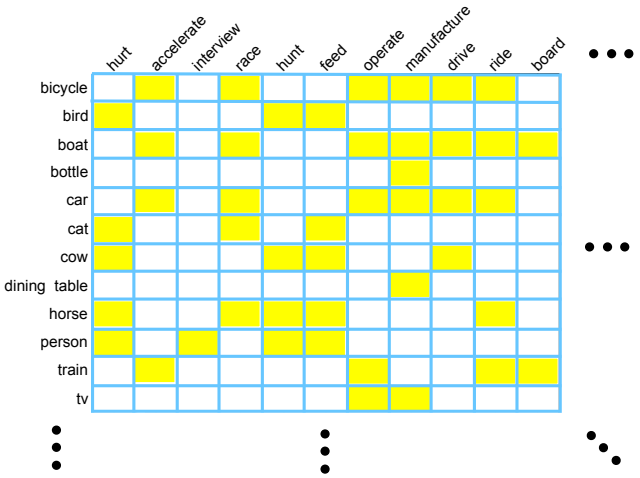


Figure 1. Mining the knowledge of semantic affordance is equivalent to filling an "affordance matrix" encoding the plausibility of each action-object pair.

In this paper, we introduce the problem of mining the knowledge of *semantic affordance*: given an action and an object, determine whether the action can be applied to the object. For example, the action of "carry" form a valid combination with "bag", but not with "skyscraper". This is equivalent to establishing connections between action concepts and object concepts, or filling an "affordance matrix" encoding the plausibility of each action-object pair (Fig. 1).

We envision that the collection of such knowledge will benefit a variety of applications. For example, the knowledge of semantic affordance provides a vocabulary for action recognition. It gives a list of valid action-object pairs, the basis for creating a large-scale image/video dataset to train and evaluate action recognition algorithms. Another application is to use the affordance knowledge as a common sense prior for generating natural language descriptions from images or videos. As demonstrated in the literature [15, 22, 31, 42, 45], such priors improves recognition of actions by eliminating implausible verb-noun combinations.

It is important to note two important distinctions of *se-*

mantic affordance in contrast with other possible representations. First, the complete knowledge of affordance is multi-faceted—it includes not only semantic relations as explored here but also spatial information such as grasp points and human poses. Complementary to prior work that primarily addressed the spatial aspect of affordances [49, 21, 46, 16, 19], we focus on the semantic aspect, leaving the spatial representation unspecified.

Second, the semantic affordance is defined in terms of *categories* of actions and objects, instead of individual "verbs" and "nouns". It is closely related, but not equivalent, to the linguistic problem of finding valid verb-noun pairs. This is because the same verb or noun can have multiple meanings (senses). For example, the meaning of the verb *draw* can be *making a drawing of* or *take liquid out of a container*. Instead, each action or object is denoted using a WordNet [30] "synset" (a set of synonyms that have the same meaning). Specifically, we aim to enrich WordNet by adding affordance edges, drawing connections between compatible verb synsets and noun synsets.

The key scientific question is: "how can we collect affordance knowledge"? We study a variety of approaches including text mining, visual mining, and collaborative filtering. Using text mining, we extract co-occurrence information of verb-noun combinations. Through visual mining, we discover whether images associated with a particular verb-noun combination are visually consistent. We also explore an interesting and surprising connection between the problem of mining semantic affordance and that of collaborative filtering: can we predict if a noun "likes" an action, just as a user likes a movie? We evaluate all approaches using ground truth affordances collected through crowdsourcing.

Our contributions are as follows: 1) we introduce the new problem of mining semantic affordances; 2) we create a benchmark dataset for affordance modeling that contains the complete ground truth for all 20 PASCAL VOC categories on 957 verb synsets;[1] 3) we explore and analyze a variety of approaches and present a number of significant insights, which open up further research for better action recognition and affordance understanding.

2. Related Work

Object Affordances There has been an emerging consensus on the benefits of modeling object affordances [49, 21, 46, 16, 19]. Gupta et al. [16] use functional constraints (human poses and motion trajectories) to aid object and action recognition. Kjellstrom et al [19] perform simultaneous action and object recognition, showing the benefits of modeling the dependency of obejcts and actions.

Another line of work seeks to discover or predict affordances on *object instances*. Zhu et al. [49] explored reasoning of affordances using a knowledge base representation. In their approach, they assume that the knowledge of semantic affordance, i.e. what we aim to mine in this paper, is already given. Yao et al. [46] discover affordances (expressed as spatial configurations of humans and objects) from weakly supervised images. Koppula et al. [21] jointly predict affordance labels and activity labels on RGB-D video segments without modeling object categories.

Our work differs from prior work on affordances mainly in that that previous research has shown the importance and benefits of using affordances, but has not addressing the issue of *collecting* the knowledge of semantic affordances.

Action Recognition Action recognition is an important problem for general image understanding and has improved dramatically over the recent years [37, 39, 43, 44]. Compared to standard datasets in object recogntion such as ImageNet [8], action recognition has been trained and evaluated with relatively small numbers of classes, e.g. 101 classes in UCF101 [40] or 487 classes in Sports-1M [18].

Due to the compositional nature of actions, there are many more action-object pairs than objects, so to advance action recognition to the next level, it is necessary to train and evaluate on a much larger number of classes. To construct such a dataset, the first question would be: what are those action-object pairs? By mining semantic affordances, we can answer this question in a systemic way.

Generating Image Descriptions This paper is also complementary to prior work that generates natural language descriptions from images and videos [15, 22, 31, 42, 45]. Previous work in this area typically uses a language model to score possible descriptions (e.g. subject-verb-object triplets). The language model is trained on text corpora and is mostly based on occurrence statistics. However, it is not clear how well linguistic scores predict semantic affordances. To the best of our knowledge, the work described in this paper provides the first analysis on this question.

Common Sense Knowledge and Attributes Mining semantic affordances is also closely connected to a wave of recent work on collecting common sense knowledge [5, 6, 13, 50, 3]. The NELL [5] project and the NEIL project [6] automatically extract structured knowledge from texts (NELL) and from images(NEIL) respectively. Another line of work leverages crowdsourcing. Freebase [3] accumulates a large number of facts through the contribution of online volunteers. Zitnick and Parikh [50] introduced visual abstraction (having humans manipulate clip art characters and objects) as a way to collect visual common sense. Fouhey and Zitnick [13] use the visual abstraction methodology to learn object dynamics. This work differs from prior work in that we focus on semantic affordances, a type of structured knowledge that was not systematically considered before.

[1] The dataset and code are available at [1]. The dataset has been extended to all 91 object categories of MS COCO [27].

Semantic affordances can also be considered as a specical type of category-level object attributes, thus connecting to work on using attributes to improve object recognition [35, 38, 47, 33, 24, 12], except that our emphasize is on collecting the attribute knowledge.

Language Understanding In computational linguistics, the line of work most closely related to ours is the one concerned with the automatic learning of selectional preferences from text. Selectional preferences (also sometime referred to as "selectional restrictions") can tell us what are the most likely arguments for a verb, e.g., "eat apple" or "dog barks". Earlier work on selectional preferences attempted to model the semantic class of the arguments by performing generalizations across semantic networks such as WordNet [30], which resulted in associations between verbs and entire semantic classes, e.g., "eat <food>". In [4] a comparison was performed of several such network-based approaches, including [36], [26], and [7], and it was found that the simpler frequency-based models can perform at par with the more advanced class-based methods. More recent data-driven work on selectional preferences attempted to identify similar arguments from large corpora [10, 32, 2], where similarity metrics are computed over vector representations of words, with the goal of identifying clusters of nouns (or classes) that can be used as arguments for a given verb. Unlike previous research, in our work we use a gold standard with extensive coverage, as well as methods that rely on visual information and collaborative filtering.

3. Crowdsourcing Semantic Affordances

The most reliable way of collecting affordance knowledge is probably crowdsourcing, i.e. asking a human whether an action can be applied to an object. At the same time, it is probably also the least scalable: it would be prohibitively expensive to exhaustively annote all action-object pairs. Nonetheless, it is feasible and necessary to obtain a subset of human annotated affordances. This serves three purposes: (1) it offers insights on how humans understand object affordances; (2) it can constitue the ground truth for evaluating automatic methods; (3) it can also be used as training data for learning-based approaches.

Selection of Objects and Actions For our crowdsourcing study, we use the 20 object categories in PASCAL VOC [11], a widely used dataset in object recognition.

The selection of action categories is not as straightforward as that of objects. Compared to objects (or noun synsets), the semantic space of actions (or verb synsets) is less well established— there is not a standard list of verb synsets that are known to be both common and "visual"— meaning that it can be depicted by images or videos. This is an important consideration because many verbs, especially those describing mental processes ("decide", "consider", "plan"), are quite hard to represent visually.

To get a list of common verb synsets, we first find out what *verbs* (without disambiguating the senses) are common. Note again the difference between verbs and verb synsets in WordNet: a verb synset is represented by one or more synonymous verbs and the same verb can appear in multiple verb synsets. We use the occurrence count of verbs in Google Syntactic N-grams dataset [28]. We start by extracting all verb-noun pairs with the *dobj* dependency, which ensures that we only get transitive verbs. We sort the extracted verbs by the total occurrence count, and create a top 1000 *verb* list. Next, we create a candidate set of *verb synsets* by taking all Wordnet verb synsets that have at least one verb in the top 1000 *verb* list.[2] This gives us a list of 2375 candidate *verb synsets*.

We set up a crowdsourcing task on Amazon Mechanical Turk (AMT) to determine the "visualness" of each candidate verb synset. In this task, we first show the definition of a verb synset with synonyms and examples, as provided by WordNet. For instance,

> **align**
> definition: place in a line or arrange so as to be parallel or straight
> synonyms: align aline line_up adjust
> example: align the car with the curb

Then we ask:

> Is it possible to tell whether someone is "align-ing" (place in a line or arrange so as to be parallel or straight) something by looking at an image or watching a video clip without sound?

Note that we repeat the definition of the verb synset in the question to make sure that it is not confused with other meanings of the same verb. A worker then chooses an answer from "Definitely yes", "Normally yes", "Maybe", "Normally no", "Definitely no", "I don't know", and "Description doesn't make sense". To further ensure the annotation quality, we also add a quiz for each verb synset to test whether the worker understands the synset definition. We detect spammers by planting a small number of gold standard questions.

For each candidate verb synset, we obtain answers from 5 different workers. For each answer, we convert it to a score ranging from 5.0 ("definitely yes") to 1.0 ("definitely no or makes no sense"). The "visualness score" of a verb synset is then the average score from 5 workers. Fig. 2 shows the distribution of scores for all candidate synsets— about 30% of the synsets have a score of 4.0 or higher.

[2] An additional criterion is that the WordNet count of the verb in the synset, a measure provided by WordNet, must be at least 3. This rules out the cases where the verb is popular but the particular verb sense is rare.

Visualness	Synset Synonyms	Definition	Example Sentence
Definitely yes	{wash, launder} {drive}	Cleanse with a cleaning agent, such as soap, and water. Operate or control a vehicle.	Wash the towels, please! Drive a car or bus.
Yes	{deliver} {switch off, cut, turn off, turn out}	Bring to a destination, make a delivery. Cause to stop operating by disengaging a switch.	Our local super market delivers. Turn off the stereo, please.
Maybe	{enjoy} {respect, honor, honour, abide by, observe}	Have for one's benefit. show respect towards.	The industry enjoyed a boom. Honor your parents!
No	{intend, destine, designate, specify} {drive}	Design or destine. Compel somebody to do something, often against his own will or judgment.	She was intended to become the director. She finally drove him to change jobs.
Definitely no /Make no sense	{wish} {come}	Make or express a wish. Come to pass, arrive, as in due course.	I wish that Christmas were over. The first success came three days later.

Table 1. Examples of verb synsets with different visualness scores.

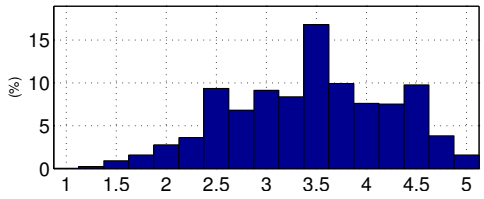


Figure 2. Histogram of visualness scores of common verb synsets.

Tab. 1 shows example synsets at different scores. Our final list of verb synsets is generated by re-ranking the candidate synsets by visualness and retaining the synsets above a cut-off visualness score (around 3.6). The cut-off score is chosen such that we have about 1,000 verb synsets.

Annotating Semantic Affordances We are now ready to annotate semantic affordances. Given an action (i.e. a verb synset) and an object (i.e. a noun synset), we ask an AMT worker whether it is possible (for a human) to perform the action on the object. Just as the visualness task, we first show the definition of the verb synset and then repeat the definition in the question, e.g.

> Is it possible to **hunt** (pursue for food or sport (as of wild animals)) a **car**?

The worker needs to choose an answer from "Definitely yes", "Normally yes", "Maybe", "Normally no", "Definitely no", "I don't know", and "Description doesn't make sense or is grammatically incorrect".

For every possible action-object pair formed by the 20 PASCAL VOC objects and the 957 visual verb synsets, we ask 5 workers to determine its plausibility. This gives a total of 19K action-object questions and 96K answers. Each answer is converted to a score from 5.0 ("Definitely yes") to 1.0 ("Definitely no or makes no sense"). The "plausibility score" of an action-object pair is then the average of 5 answers.

Analysis Fig. 3 shows the distribution of plausibility scores for all action-object pairs. On average, around 24% of action-object pairs have scores 4.0 or higher. Tab. 2 shows examples of action-object pairs with different plausibility scores.

How do the semantic affordances differ between objects? It has long been hypothesized that object catgories

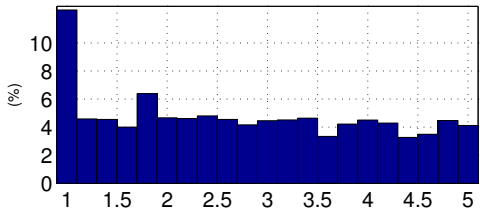


Figure 3. Distribution of human annotated plausibility scores for all action-object pairs

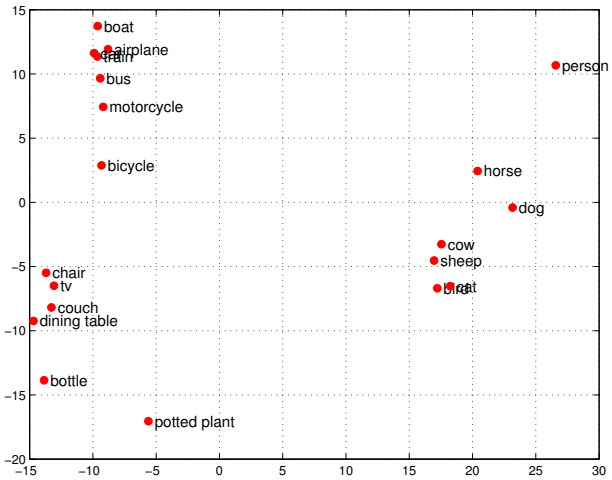


Figure 4. Visualizing 20 PASCAL VOC object classes in the semantic affordance space.

are formed based on functionality [14]. Our exhaustive annotations provide an opportunity to confirm this hypothesis. Each object has a 957 dimensional "affordance vector", where each dimension is the plausibility score with an action. We use PCA to project the affordance vectors to a 2-dimensional space and plot the coordinates of the object classes in Fig. 4. It is notable that the object classes form clusters that align well with a category-based semantic hierarchy—we can clearly see one cluster for vehicles, one for animals, and one for furniture. This validates the functional view of semantic categories.

What affordances best distinguish the different object classes? We sort the entries of the first two principal components by their absolute values, and look at the associated verb synstes of to those entries. Fig. 5 shows the plausibility scores of these actions on several objects in the PASCAL-20 set. This suggests that affordances are indeed very discrim-

Plausibility	Action		Object	
	Synset Synonyms	Definition	Synset Synonyms	Definition
Definitely yes	{race, run}	Compete in a race.	{car, auto, automobile, machine, motorcar}	A motor vehicle with four wheels; usually propelled by an internal combustion engine.
	{feed, eat}	Take in food; used of animals only.	{dog, domestic dog, Canis familiaris}	A member of the genus Canis that has been domesticated by man since prehistoric times.
Yes	{repel, repulse, fight off, rebuff, drive back}	Force or drive back.	{bear}	Massive plantigrade carnivorous or omnivorous mammals with long shaggy coats and strong claws.
	{turn}	Cause to move around or rotate.	{sofa, couch, lounge}	An upholstered seat for more than one person.
Maybe	{compress, constrict, squeeze, compact, contract, press}	Squeeze or press together.	{bottle}	A glass or plastic vessel used for storing drinks or other liquids; typically cylindrical without handles.
	{repair, mend, fix, bushel, doctor furbish up, restore, touch on}	Restore by replacing a part or putting together what is torn or broken.	{wineglass}	A glass that has a stem and in which wine is served.
No	{capture, catch}	Capture as if by hunting, snaring or trapping.	{chair}	A seat for one person, with a support for the back.
	{ignite, light}	Cause to start burning; subject to fire or great heat.	{knife}	Edge tool used as a cutting instrument; has a pointed blade with a sharp edge and a handle
Definitely no /Make no sense	{cultivate, crop, work}	Prepare for crops.	{person, individual, someone somebody, mortal, soul}	A human being.
	{wear, bear}	Have on one's person.	{airplane, aeroplane, plane}	An aircraft that has a fixed wing and is powered by propellers or jets.

Table 2. Examples of action-object pairs with different average plausibility scores.

inative attributes for object categories.

4. Mining Semantic Affordances

In this section we study approaches that mine semantic affordances. We pose the question: to what extent can we automatically extract such information?

4.1. Mining from Texts

We first explore the possibility of mining from texts. To determine the plausibility of an action-object pair, we consider the following signals from texts:

- **N-Gram Frequency**. We count the frequency of the verb-noun pair in Google Syntactic N-Grams. This is the basis of many language models used in the literature for generating descriptions from images [15, 22, 31, 42, 45].
- **Latent Semantic Analysis (LSA)**. LSA [25] is a widely used method to convert words to semantic vectors, which can then be used to compute the similarity of two words. LSA essentially factorizes the word-document matrix. Words that tend to co-occur in the same document would get mapped to similar vectors. Given a verb and a noun, we compute their cosine similarity as a proxy for plausibility of affordance. To train the model, we use the Europarl parallel corpus [20] by segmenting the corpus into sentences and training for 2 cycles.
- **Word2Vec** [29] is the current state of the art method for embedding words into semantic vectors. At its core is a deep neural network trained to predict a word based on its surrouding context. We use the Continuous Bag-of-Words architecture and train on the same corpus as LSA. We set the dimensionality to 300, window size to 5, and train for 15 iterations. Similar to LSA, we compute the similarity between the verb and the noun as a signal for affordance.

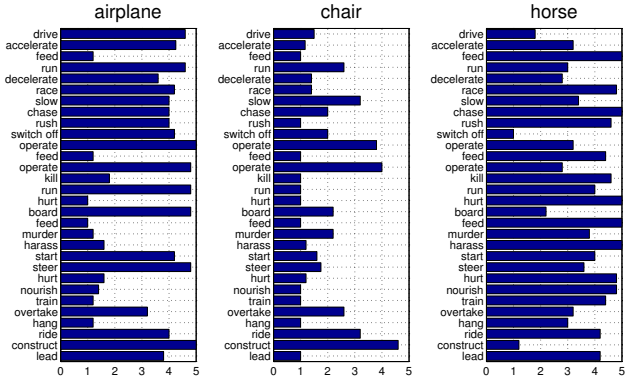


Figure 5. Plausibility scores on the verb synsets that have high responses in the first two principal components.

Since all these methods operate at the word level instead of the sense level, in cases of multiple verbs for a synset, we use the leading verb in the synset, the most representative verb among the synonyms as specified by WordNet.

Evaluation How well do these signals from texts predict semantic affordances? To evaluate, we first binarize the plausibility scores using a threshold 4.0 (average answer is "normally yes" or above) [3]. Thus the problem of predicting affordance becomes a binary classification problem: given an object and an action, predicting the pair to be plausible or not. Following the tradition of PASCAL VOC, we evaluate each object separately and then compute the average. For each object, we rank the verb synsets using one of the textual signals, plot a precision recall curve, and compute the average precision (AP). We also evaluate mAP, the mean average precision over all objects.

Results Fig. 6 shows the precision recall curves for each signal. Tab. 4 presents the (mean) average precision. The results show that the textual signals are indeed predictive

[3]This is the threshold used throughout the paper. We have also run all experiments with the threshold 3.0, which produce similar results and do not affect the conclusions we draw.

of semantic affordance, but they are very far from perfect. They can accurately retrieve a small number of affordances but the precision drops quickly with a higher recall. Somewhat surprisingly, the simplest method, Google N-Gram, outperforms the more sophisticated LSA and Word2Vec. This is likely because LSA and Word2Vec consider larger context windows, which may introduce spurious associations between verbs and nouns, whereas Google N-Gram only considers verbs and nouns with direct dependencies.

Tab. 3 shows success and failure cases of the Google N-Gram signal. We see that the false positives can be attributed to two cases: 1) no disambiguation of the verb (e.g. "pass a bottle" where "pass" means "go across or through"), and 2) failure in parsing the semantic dependency between the verb and the noun (e.g. "feed bus" has a high count likely because the original texts were about "twitter feed on bus schedule"). The false negatives reveal a more fundamental limitation with the text based signals: what if something has not appeared in the corpus? For example, "photograph an airplane" has a count zero in the Google N-Gram dataset, but it is a perfectly valid action-object pair.

4.2. Mining from Images

We now investigate mining from images, another possible source of affordance knowledge. The idea is that we can use the verb-noun pair representing the action-object affordance to query an image search engine. Search engines can rely on historical user click data to identify the images that match the query. Thus the top images returned by a search engine may be assumed to be correct. Under this assumption, if the affordance exists, the top returned images should be more visually coherent. If the affordance does not exist, the returned images would be more random.

Similar ideas have been explored by prior work (e.g. [9, 6]). For example, the LEVAN system developed by Divvala et al. [9] queries Google Image Search to discover new visual concepts. It verifies that the concept is visually valid by checking the visual consistency of the returned images. Following their approach, we train a binary classifier to differentiate the top returned images against a set of random background images. The cross-validation accuracy of this classifier can then be used as a consistency measure for the returned images. In particular, we train an SVM classifier using features extracted by Alex's Net [23] implemented in CAFFE [17]. We also use Google Image Search as the source of images.

Results The question is how well this approach would work for predicting semantic affordances. Evaluating the visual consistency signal the same way as the individual textual signals in Sec. 4, we plot the precision recall curves (Fig. 6) and present the average precision (Tab. 4). The results indicate that although decent precision can be achieved with a very low recall, the precision recurve curve is very poor—in fact, not better than random most of the times. Thus the visual signals are much worse than textual signals.

Fig. 7 shows sample image search results and the corresponding accuracy of the learned classifier (i.e. visual consistency). Inspecting these results reveals several sources of errors. False positives arise when Google Image Search can return images that are irrelevant to the query but are highly visually uniform due to accidental textual match, e.g. the queries "wear bicycle" and "transport chair" return visually uniform images, but the content of the images are very different from the underlying concepts of the queries. False negatives occur when the search engine either fails to return enough relevant images (e.g. "manufacture chair" in Fig. 7), or when many returned images are relevant but too visually diverse to learn a classifier even with the current state of the art feature representation (e.g. "wash bicycle" in Fig. 7).

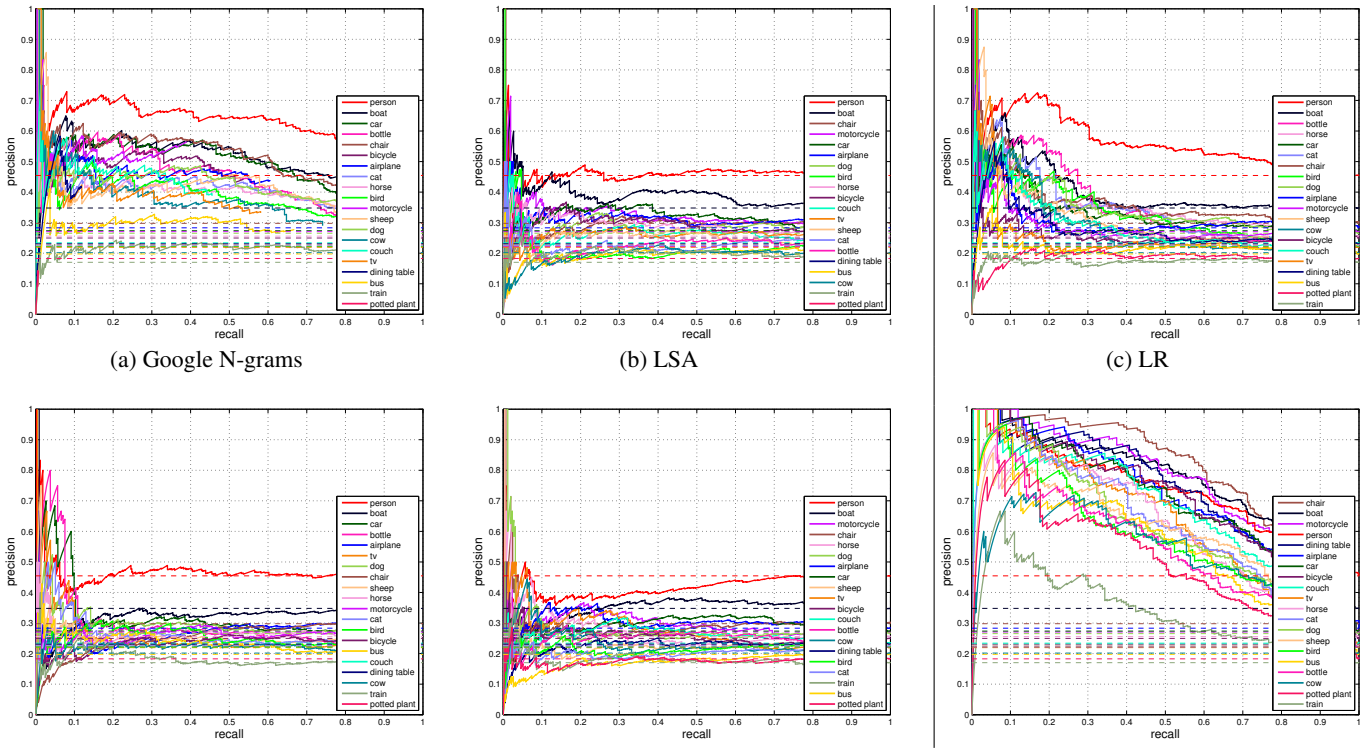
4.3. Collaborative Filtering

So far we have explored approaches that use signals on *individual* action-object pairs. However, performing PCA on the human annotated affordances (Fig. 4) suggests that the space of affordance vectors is lower dimensional and "smooth". This observation leads to an interesting connection to the problem of collaborative filtering [41] (or matrix completion): suppose we already observe the ratings of some users on some movies, can we predict the rest of the ratings? Here we just need to replace "user" with object and "movie" with action. This connection opens up the possibility of *extrapolation*, i.e. inferring new affordances based on existing ones.

Some collaborative filtering methods admit "side information", attributes or features about the users or movies in addition to the observed ratings. Including side information allows us to handle the "cold start" scenario, where for certain users (or movies) we do not have any observed ratings. Without loss of generality, side information can be expressed as similarities (kernels) between users or between movies.

We employ Kernelized Probabilistic Matrix Factorization (KPMF) [48], a state of the art collaborative filtering method that allows similarity based side information. Essentially, for N objects and M actions, the method factorizes the $N \times M$ affordance matrix into the product of a latent $N \times D$ matrix and a latent $D \times M$ matrix. There is an additional constraint: if we treat each row (column) of the latent $N \times D$ ($D \times M$) matrix as a new representation of the corresponding object (action), then similarities based on this new representation should agree with the known similarities provided as side information.

Evaluation Following the evaluation setting in Sec. 4.1, we evaluate each object class separately. Given an object class, we assume that none of its affordances are observed but the ground truth affordances for all the other 19 ob-



(a) Google N-grams (b) LSA (c) LR

(d) Word2Vec (e) V Consistency (f) KPMF

Figure 6. PR curves. Each row corresponds to a different score threshold when setting the ground truth. Each column corresponds to different baselines: (a) occurence count from Google syntatic N-grams, (b) LSA, (c) Word2Vec, (d) Visual Consistency, (e) logistic regression on (a),(b),(c),(d), and (f) Collaborative filtering (KPMF). Dash lines represent chances.

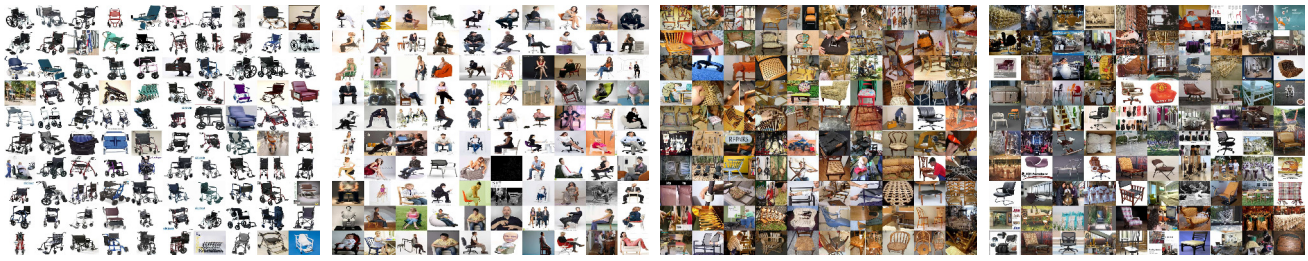
Google N-Gram	Action		Object	
	Synset Synonyms	Definition	Synset Synonyms	Definition
True positives	{fly, aviate, pilot}	Operate an airplane.	{airplane, aeroplane, plane}	An aircraft that has a fixed wing and is powered by propellers or jets.
	{draw}	Represent by making a drawing of, as with a pencil, chalk, etc. on a surface.	{person, individual, someone, somebody, mortal, soul}	A human being.
	{pass, hand, reach, pass on, turn over, give}	Place into the hands or custody of.	{bottle}	A glass or plastic vessel used for storing drinks or other liquids; typically cylindrical without handles.
False positives	{fly}	Transport by aeroplane.	{airplane, aeroplane, plane}	An aircraft that has a fixed wing and is powered by propellers or jets.
	{draw, take out}	Take liquid out of a container or well.	{person, individual, someone, somebody, mortal, soul}	A human being.
	{pass,go through,go across}	Go across or through.	{bottle}	A glass or plastic vessel used for storing drinks or other liquids; typically cylindrical without handles.
	{feed, give}	Give food to.	{bus, autobus, coach, charabanc, double-decker}	A vehicle carrying many passengers; used for public transport.
False negatives	{photograph, snap, shoot}	Record on photographic film.	{airplane, aeroplane, plane}	An aircraft that has a fixed wing and is powered by propellers or jets.
	{award, present}	Give, especially as an honor or reward.	{person, individual, someone, somebody, mortal, soul}	A human being.

Table 3. Examples of success and failure cases for Google N-Gram, the best performing text-based signal

	aero	bike	bird	boat	bottle	bus	car	cat	chair	couch	cow	table	dog	horse	mbike	person	plant	sheep	train	tv	mAP
Random	28.3	25.0	22.2	34.8	22.2	19.7	27.4	22.6	29.8	23.4	20.2	23.1	26.5	25.6	27.1	45.5	18.3	22.8	17.0	21.9	25.2
G N-Grams [28]	44.1	44.4	41.4	53.9	47.9	27.5	50.0	43.3	47.8	36.7	39.5	29.5	40.6	42.2	41.2	65.3	19.5	40.6	21.6	36.7	40.7
LSA [25]	31.5	28.8	29.0	39.9	24.4	21.2	31.7	25.3	35.5	27.8	20.7	23.1	30.1	28.9	34.0	47.4	18.3	26.6	19.4	27.4	28.5
Word2Vec [29]	31.4	24.8	25.5	40.0	31.4	24.3	33.0	26.8	29.0	23.4	22.0	23.1	30.2	28.2	27.6	50.5	18.3	28.9	20.1	30.7	28.5
V Consistency	33.2	28.2	23.6	38.5	26.8	20.1	31.7	22.7	36.8	28.2	25.2	24.3	33.9	34.2	36.9	48.2	19.8	30.6	21.4	29.0	29.7
LR	35.6	33.3	38.5	45.2	39.7	23.6	39.2	38.7	38.7	31.8	34.1	30.1	36.5	39.4	35.5	60.0	20.0	34.2	18.5	30.7	35.2
NN	55.6	52.7	49.7	63.0	49.5	44.6	61.9	50.1	66.3	58.1	47.5	60.2	56.8	55.3	61.6	57.2	28.2	51.7	**41.2**	51.0	53.1
KPMF [48]	**71.1**	**69.9**	**58.3**	**76.3**	**56.4**	**56.8**	**70.2**	**62.0**	**78.1**	**67.1**	**53.6**	**71.5**	**61.8**	**62.1**	**75.2**	**73.6**	**50.5**	**59.7**	36.2	**63.3**	**63.7**

Table 4. Per object average precision (AP) and mean average precision (mAP) for a variety of automatic mining methods.

wear + bicycle, 90.00 % | ride + bicycle, 87.50 % | lock + bicycle, 84.00 % | wash + bicycle, 63.50 %



transport + chair, 93.50 % | sit + chair, 92.50 % | repair + chair, 81.50 % | manufacture + chair, 73.00 %

Figure 7. Query keywords, the top 100 images return by Google Image Search, and visual consistency (cross validation accuracy).

ject classes are available. We define side information as the WordNet similarities (e.g. PATH, LCH, WUP [34]) between two objects. We then run KPMF to predict plausibility scores for the unobserved entries using the observed ground truths. For each of the 20 object classes, we repeat this process and report a precision recall curve and an average precision (AP).

We compare KPMF with two baselines. First, we predict the affordances for each object class by simply transferring the affordance labels from the most similar object class among the other 19 (NN). Second, we learn a logistic regression (LR) classifier that linearly combines the textual and visual signals to predict the plausibilities. For each object class, the weights of the classifier is learned using the ground truth plausibility on the other 19 objects.

Results The results are reported in Tab. 4 and Fig. 6. Across all object classes, collaborative filtering (KPMF) outperforms textual and visual signals by a very large margin, 63.7 mAP versus 40.7. Besides, KPMF also outperforms LR and NN, suggesting that both side information and matrix factorization are essential. Interestingly, the logistic regression classifier performs worse than Google N-Gram, suggesting that the learned weights do not generalize across classes. These results confirm that collaborative filtering can indeed exploit the low rank structure of the affordance matrix and generalize to new classes using side information, *which surprisingly turns out to be much more effective than mining from texts and images.*

5. Discussions and Future Work

Our study has led to a number of interesting findings: 1) Human annotated ground truth affordances have low dimensional structure that reveals a good alignment of functionalities and categories; 2) Language models based on co-occurrence statistics have substantial limitations in predicting affordances due to the difficulty of sense disambiguation and inevitable data sparsity; 3) Visual models are significantly weaker than language models in predicting affordances. 4) Collaborative filtering can effectively exploit the low rank structure of affordances.

These findings suggest a plausible bootstrapping sequence towards better affordance knowledge and action recognition. We first use crowdsourcing and collaborative filtering to collect an initial set of high quality affordances, which can in turn help improve visual and language models. We then use the improved visual and language models to predict more affordances and form a virtuous cycle. This human-collaborating strategy can also lead to possible cost reduction in future data collection.

Since we are studying a new problem, we start with the simplest possible semantic affordances: action-object pairs (transitive verb + noun). Note that semantic affordances can also be described by intransitive verbs, e.g. cutting with a knife, or more complex verb-noun relationships, e.g. action-object-instrument triplets.

Finally, reasoning about affordances—determining whether an action can be performed on an object—is in itself a meaningful and challenging problem in AI, as it requires multimodal common sense reasoning involving both vision and language. Our study introduces this problem, establishes the first benchmark, and presents a number of new insights. We believe that this work will open up a rich space for further exploration.

References

[1] http://www.umich.edu/~ywchao/semantic_affordance/. 2

[2] S. Bergsma, D. Lin, and R. Goebel. Discriminative learning of selectional preference from unlabeled text. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, 2008. 3

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proc. of the 2008 ACM SIGMOD International Conf. on Management of Data*, 2008. 2

[4] C. Brockmann and M. Lapata. Evaluating and combining approaches to selectional preference acquisition. In *Proc. tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, 2003. 3

[5] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *Proc. AAAI Conf. on Artificial Intelligence*, 2010. 2

[6] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting visual knowledge from web data. In *Proc. IEEE International Conf. on Computer Vision*, 2013. 2, 6

[7] S. Clark and D. Weir. Class-based probability estimation using a semantic hierarchy. In *Proc. Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001. 3

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. 2

[9] S. Divvala, A. Farhadi, and C. Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. 6

[10] K. Erk. A simple, similarity-based model for selectional preferences. In *Proc. Association of Computational Linguistics*, 2007. 3

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 3

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. 3

[13] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. 2

[14] J. J. Gibson. *The ecological approach to visual perception*. Psychology Press, 2013. 4

[15] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proc. IEEE International Conf. on Computer Vision*, 2013. 1, 2, 5

[16] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009. 2

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM International Conf. on Multimedia*, 2014. 6

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classication with convolutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. 2

[19] H. Kjellström, J. Romero, and D. Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011. 2

[20] P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, 2005. 5

[21] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 2

[22] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. In *Proc. 27th AAAI Conf. on Artificial Intelligence*, 2013. 1, 2, 5

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012. 6

[24] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009. 3

[25] T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997. 5, 7

[26] H. Li and N. Abe. Generalizing case frames using a thesaurus and the mdl principle. *Computational linguistics*, 24(2):217–244, 1998. 3

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. Zitnick. Microsoft coco: Common objects in context. In *Proc. of the European Conf. on Computer Vision*. 2014. 2

[28] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov. Syntactic annotations for the google books ngram corpus. In *Proc. ACL 2012 System Demonstrations*, 2012. 3, 7

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013. 5, 7

[30] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2, 3

[31] T. S. Motwani and R. J. Mooney. Improving video activity recognition using object recognition and text mining. In *Proc. 20th European Conf. on Artificial Intelligence*, 2012. 1, 2, 5

[32] P. Pantel, R. Bhagat, B. Coppola, T. Chklovski, and E. Hovy. ISP: Learning inferential selectional preferences. In *Human Language Technologies 2007: The Conf. of the North American Chapter of the Association for Computational Linguistics; Proc. Main Conf.*, 2007. 3

[33] D. Parikh and K. Grauman. Relative attributes. In *Proc. IEEE International Conf. on Computer Vision*, 2011. 3

[34] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity: Measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004*, 2004. 8

[35] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. 3

[36] P. Resnik. Selectional preference and sense disambiguation. In *Proc. ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, 1997. 3

[37] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012. 2

[38] A. Shrivastava, S. Singh, and A. Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *Proc. European Conf. on Computer Vision*. 2012. 3

[39] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27*. 2014. 2

[40] k. Soomro, A. Roshan Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. In *CRCV-TR-12-01*, 2012. 2

[41] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009:4, 2009. 6

[42] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proc. 25th International Conf. on Computational Linguistics (COLING)*, 2014. 1, 2, 5

[43] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. IEEE International Conf. on Computer Vision*, 2013. 2

[44] H. Wang and C. Schmid. Lear-inrea submission for the thumos workshop. In *ICCV Workshop on Action Recognition with a Large Number of Classes*, 2013. 2

[45] Y. Yang, C. L. Teo, H. Daumé, III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, 2011. 1, 2, 5

[46] B. Yao, J. Ma, and L. Fei-Fei. Discovering object functionality. In *Proc. IEEE International Conf. on Computer Vision*, 2013. 2

[47] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. 3

[48] T. Zhou, H. Shan, A. Banerjee, and G. Sapiro. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *Proc. 2012 SIAM International Conf. on Data Mining*, 2012. 6, 7

[49] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *Proc. European Conf. on Computer Vision*, 2014. 2

[50] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. 2