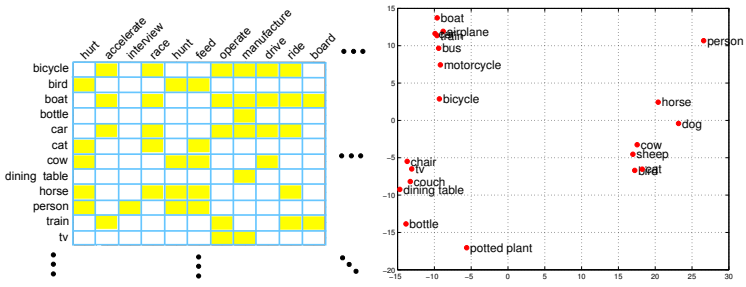
Mining Semantic Affordances of Visual Object Categories

Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng
Computer Science & Engineering, University of Michigan, Ann Arbor



(a) (b)

Figure 1: (a) "Affordance matrix" encoding the plausibility of each action-object pair. (b) 20 PASCAL VOC object classes in the semantic affordance space.

Affordances are fundamental attributes of objects. Affordances reveal the functionalities of objects and the possible actions that can be performed on them. We can "hug" a dog, but not an ant. We can "turn on" a tv, but not a bottle. Acquiring such knowledge is crucial for recognizing human activities in visual data and for robots to interact with the world. The key question is: given an object, can an action be performed on it? While this might seem obvious to a human, there is no automated system that can readily answer this question and there is no knowledge base that provides comprehensive knowledge of object affordances.

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). The key scientific question is: "how can we collect affordance knowledge"? We first introduce a new benchmark with crowdsourced ground truth affordances on 20 PASCAL VOC object classes and 957 action classes. We then study a variety of approaches including 1) text mining, 2) visual mining, and 3) collaborative filtering. We quantitatively evaluate all approaches using ground truth affordances collected through crowdsourcing.

For our crowdsourcing study, we ask human annotators to label whether an action-object pair is a valid combination. We use the 20 object categories in PASCAL VOC [2]. We design experiments to obtain a list of action categories that are both common and "visual". Our list contains 957 action categories extracted from the verb synsets on Wordnet [6] that has 1) a member verb that frequently occurs in text corpora, and 2) high "visualness score" determined by human labelers. Given the list of actions and objects, we set up a crowdsourcing task on Amazon Mechanical Turk (AMT). We ask crowd workers whether it is possible (for a human) to perform a given action on a given object. For instance,

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

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

What is the distribution of 20 PASCAL object classes in their affordance space? We answer this by analyzing the human annotated affordances. 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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

	Rand	N-Grams [4]	LSA [3]	Word2Vec [5]	V Consistency	LR	NN	KPMF [7]
mAP	25.2	40.7	28.5	28.5	29.7	35.2	53.1	**63.7**

Table 1: Mean average precision (mAP) for a automatic mining methods

classes in Fig. 4. It is notable that the object classes form clusters that align well with a category-based semantic hierarchy.

Next, we study to what extent we can automatically extract the affordance information. We investigate three different mining approaches:

1. Mining from Texts We determine the plausibility of an action-object pair by considering the following signals from texts

- Frequency of the verb-noun pair in **Google Syntactic N-Grams** [4].
- Similarity obtained by **Latent Semantic Indexing (LSA)** [3].
- Similarity obtained by **Word2Vec** [5].

2. Mining from Images We use the verb-noun pair representing the action-object affordance to query an image search engine. Assuming top images returned by a search engine may be correct, if the affordance exists, the top returned images should be more visually coherent. Otherwise, the returned images would be more random. We measure the visual consistency by the cross-validation accuracy of a classifier trained to differentiate the top returned images against a set of random background images

3. Collaborative Filtering We ask the question that, suppose we already observe the affordance labels of some object on some action, can we predict the rest of the ratings? We investigate Kernelized Probabilistic Matrix Factorization (KPMF) [7], a state of the art matrix factorization based method that exploits side informations. Given 19 object classes with fully-observed affordances, we use KPMF to predict plausibility scores for the unobserved object class.

We evaluate the affordance prediction as 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. Tab. 1 presents the mean average precision (mAP) over all 20 object categories for each approach. Our results show that collaborative filtering (KPMF) significantly outperforms language and visual models.

In conclusion, our study introduces a new problem, establishes the first benchmark, and presents a number of new insights. We have made our dataset and code publicly available [1].

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

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

[3] 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.

[4] 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.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

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

[7] Tinghui Z., Hanhuai S., Arindam B., and Guillermo S. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *SDM*, 2012.