Open-vocabulary Object Retrieval

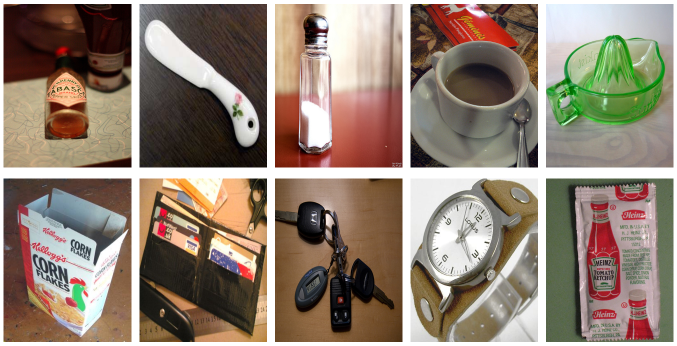
Sergio Guadarrama*, Erik Rodner†, Kate Saenko‡, Ning Zhang*, Ryan Farrell*†, Jeff Donahue* and Trevor Darrell*†
*EECS, University of California at Berkeley
{sguada, nzhang, farrell, jdonahue, trevor}@eecs.berkeley.edu
†International Computer Science Institute (ICSI)
erik@icsi.berkeley.edu
‡University of Massachussetts, Lowell
saenko@cs.uml.edu

***Abstract*—In this paper, we address the problem of retrieving objects based on open-vocabulary natural language queries: Given a phrase describing a specific object, e.g., "the corn flakes box", the task is to find the best match in a set of images containing candidate objects. When naming objects, humans tend to use natural language with rich semantics, including basic-level categories, fine-grained categories, and instance-level concepts such as brand names. Existing approaches to large-scale object recognition fail in this scenario, as they expect queries that map directly to a fixed set of pre-trained visual categories, e.g. ImageNet synset tags. We address this limitation by introducing a novel object retrieval method. Given a candidate object image, we first map it to a set of words that are likely to describe it, using several learned image-to-text projections. We also propose a method for handling open-vocabularies, i.e., words not contained in the training data. We then compare the natural language query to the sets of words predicted for each candidate and select the best match. Our method can combine category- and instance-level semantics in a common representation. We present extensive experimental results on several datasets using both instance-level and category-level matching and show that our approach can accurately retrieve objects based on extremely varied open-vocabulary queries. The source code of our approach will be publicly available together with pre-trained models at http://openvoc.berkeleyvision.org and could be directly used for robotics applications.**

I. Introduction

Visual recognition can semantically ground interaction in a physical environment: when we want a robot to fetch us an object, we may prefer to simply describe it in words, rather than use a precise location reference. But what label to use? ImageNet synsets? LabelMe tags? Should we refer to its fine-grained category, brand-name, or describe the specific instance? Or maybe use product identifiers from an online merchant? Clearly, we need visual recognition methods which accommodate the full range of natural language and situation-specific lexical biases when resolving users' references to objects [21].

Large-vocabulary object recognition has recently made significant advances, spurred by the emergence of dictionary-scale datasets [15, 32]. Dramatic progress has been made on category-level recognition [28], where each image is classified as one or more basic-level nouns, e.g. *bird, car, bottle*, and on fine-grained recognition of hundreds of specific species or sub-categories, e.g. *sparrow, Prius, Coke bottle* [7]. Addressing the *open-vocabulary* problem that arises when natural language strings are the label space requires recognizing potentially millions of separate categories [4]. In the robotic perception setting, detecting object instances often requires good RGB-D models of the actual objects to be recognized [50, 56], and therefore do not scale well to more than a hundred objects.



"Please, select the ________________________________ "

a) *bottle which is lying down / pepper sauce bottle please / Tabasco / bottle of Tabasco sauce / Tabasco brand sauce / sauce here / Tabasco pepper sauce / red Tabasco sauce / small glass bottle*

b) *empty corn flakes box / white box of cereal / corn flakes / corn flakes pack / corn-flakes packet / Kellogg's corn flakes*

Fig. 1. An open-vocabulary object retrieval task. A user describes an object of interest using natural language, and the task is to select the correct object in a set or scene. A mixture of instance-level and category-level references are typically provided by users when naturally referring to objects. The phrases listed in (a) and (b) were produced by users referring to the first images in the top and bottom rows, respectively.

In this paper, we address the task of open-vocabulary object retrieval using descriptive natural language phrases by combining category and instance level recognition. Given a phrase describing a *specific* object, our goal is to retrieve the best match from a set of unlabeled image regions containing candidate objects. As illustrated above, this problem arises in situated human-machine interaction: users interacting with situated robots often refer to objects of interest in a physical environment using natural language utterances. For example,

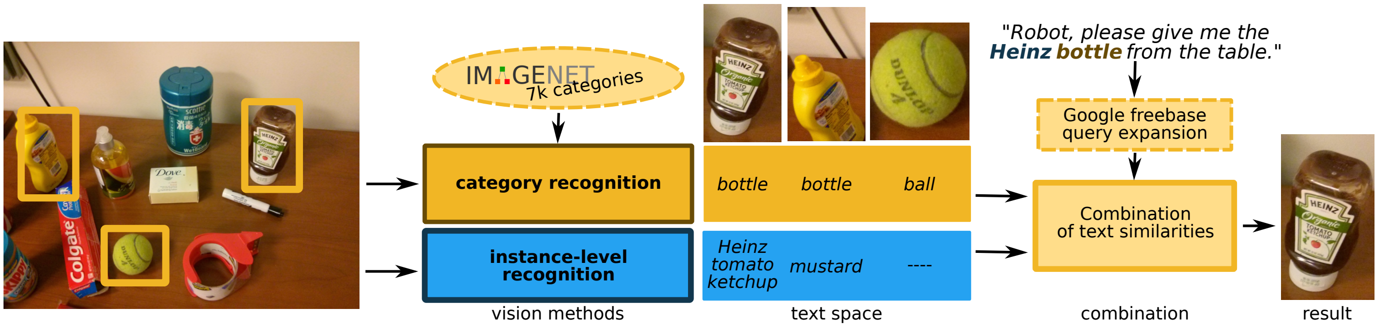

IMAGENET
7k categories
category recognition
instance-level recognition
vision methods
bottle
bottle
ball
Heinz tomato ketchup
mustard
----
text space
"Robot, please give me the Heinz bottle from the table."
Google freebase query expansion
Combination of text similarities
combination
result

Fig. 2. Multiple image projections project a candidate image windows into a semantic text space, by employing text associated with synset definitions and text associated with matched images. The user's query is also projected into this space, and the closest match is returned.

a user might ask a robot to find and bring "empty corn flakes box" (Fig. 1). In such scenarios, humans rarely name an object with a single basic-level noun (e.g., "box"). Rather, they use a rich and varied set of words including attributes (e.g., "white"), brand names (e.g., "Kellogg's corn flakes"), and related concepts (e.g., "cereal"). In fact, in our experiments, human subjects mentioned the main category noun in only 60% percent of descriptions. Further, even when they do use a basic-level category—"box", in this example—the precision of retrieval using that category may be much lower than that obtained by directly matching an image instance, e.g., matching the logo likely associated with "Corn Flakes" here.

While *generating* descriptive nouns, attributes, and/or phrases from an image is a topic of recent interest [30, 16, 41], the challenge of *retrieving* an image or object from within a scene using natural language has had less attention (but see [24, 52, 51]). We frame the problem as one of content-based image retrieval, but searching a relatively small set of potential distractors rather than sifting through a large image corpus. Our method is inspired in part by recent methods which employ image-to-text projections for large-scale recognition [55, 49, 20]. Instead of mapping an image to a set of category labels, we map it to a sparse distribution over an open-vocabulary text space. This allows us to predict words related to the specific object image at the level of instances, fine-grained categories, or categories at any level of the semantic hierarchy, plus other words related to the object.

We propose an approach that leverages the semantics of categories, subcategories, and instance-level semantics, combining them in a common representation space defined by word distributions (Fig. 2). Candidate images are projected into the common space via a set of image-to-text projections. We propose a combination scheme that ranks the candidate images in a cascaded fashion, using projections in the order of highest to lowest expected precision, until a confident match is found.

Our framework incorporates a variety of category classification and instance matching methods to define image-to-text projections. At the category level, we consider both a conventional bank of linear SVMs on sparse coded local features [15] and a deep convolutional model trained on ImageNet [28], and define the projection to a text space based on the text that defines the associated synset. At the instance level, we use large-scale image search engines [3, 2] to index product images and other images available on the web to find matching web pages, and take the text from those pages. We expand query terms using the Freebase API [1], so that semantically related terms are included such that the chance of a match is increased for each projection (at some cost to precision but improving the coverage, as our experiments reveal).

We evaluate our methods on a subset of ImageNet test data corresponding to categories which are relatively dense with household product objects, and on new images collected in a robotics laboratory. We show each image to human annotators and ask them to provide a description for a robot to retrieve it from a room in their home. We then evaluate our method's ability to find the correct object in simulated scenes of varying complexity. To the best of our knowledge, ours is the first method proposed to fuse instance-level and category-level semantics into a common representation to solve open-vocabulary reference tasks. Our results show that our sequential cascade approach outperforms a variety of baselines, including existing category-level classifiers or instance-level matching alone, or a baseline formed by matching images returned by a text-based image search as proposed in [5].

II. Related Work

Object recognition and content-based image retrieval each have a rich history and a full review of either is beyond the scope of this summary; most relevant to this paper perhaps is the relatively recent work on large scale categorization. Many approaches are trained on ImageNet [15], and output a single category label [16, 28], while others try to generate proper natural language descriptions [18]. Recent efforts have investigated increasingly fine-grained recognition approaches [19, 42], predicting e.g., the specific breed of dogs, or the model and year of a car. Instance-level recognition has a long history in computer vision [34, 37, 43, 47], and has been successfully deployed in industry for a variety of products (e.g., Google Goggles).

In robotics perception it has shown very good results in instance recognition when training from RGB-D data of the objects of interest [50, 56]. However these approaches generally require precise RGB-D models of the objects to be recognized and therefore do not scale beyond a predifined set of objects. Recent work [27] shows how to learn a logical model of object relations to allow for object retrieval. Therefore this paper is closely related to ours when it comes to the application scenario. Despite the attempt of training one-vs-all classifiers for hundreds of thousands of labels [14], no fixed vocabulary of nouns is sufficient to handle open-vocabulary queries, which involve arbitrary labels at all levels of semantics, from generic to extremely fine-grained to attribute-level. More importantly, the amount of supervised data required for each of these constituent problems presents a major barrier to enabling arbitrary vocabularies. Our proposal can be seen as complementary to previous approaches to object recognition in robotics [50, 56], in that it handles novel objects and out-of-vocabulary labels.

Earlier work has focused on modeling co-occurrences between image regions and tags [6, 8], focusing on scenes where the correspondence between image regions and tags is unknown. Hwang and Grauman [25] propose to extract features from a given ordered list of tags and use them to estimate the size and location of an object in an image. In contrast, we don't assume that we have paired text and images for training, we use them only for validation and testing.

A related line of work embeds corresponding text and image pairs in a common space, such that both the image and the text end up at nearby locations (Canonical Correlation Analysis, Kernelized Canonical Correlation Analysis, *etc*.). The major limitation of such embeddings is that they in general do not include both category- and instance-level labels, and require training images paired with text. In our case, such data is available for some objects through search-by-image engines, but not for all. The work of [46] proposes a general framework for supervised embeddings; this and related efforts could be profitably applied to enhance our representation, assuming one could obtain a lot of images paired with text, but is not necessary to obtain the results we report in this paper.

Caption-based retrieval methods (Kulkarni *et al.* [29], *etc*.), map images to text captions, but focus on scenes rather than objects and category-level rather than instance-level information. Moreover, they rely on captioned images for training data. Several web-scale image tagging methods consider applications to tag-to-image search, or image retrieval using text-based queries [23, 26, 33, 35]. Most have been limited to one-word queries and category-level tags, e.g., *pool*, and cannot handle phrase queries or queries that may contain instance-level tags, e.g., *Froot Loops cereal.*

Instance recognition methods try to find a set of relevant images given a query image. For example, [22] proposed to use category-level labels to improve instance-level image retrieval and use a joint subspace and classifier learning to learn a projection in a reduced space using category labels.

Another line of work within image retrival is [5, 11], where authors try to find a set of relevant images in a dataset given a short text query. [5] proposes using Google Image Search to find candidate image queries and then use those to rank the images in the dataset. However, they use a very restricted set of queries, and a small dataset. Nonetheless, we evaluate this method as a baseline to compare against our method, as reported below. In [11] authors also propose to use Google Image Search to train an object classifer on-the-fly and use it to rank the images in the dataset. In this case authors use a large set of distractors from ImageNet, but only evaluate their approach on Pascal VOC 2007, where there are just 20 classes. It is not clear if their approach could be used for thousands of classes with open-vocabulary queries like ours.

In the video retrieval setting, several papers addressed the problems created by using natural language in the queries [31, 38, 45, 48, 39]; more recently [13] addressed the problem of zero-shot video retrieval using content and concepts.

III. Open-vocabulary object retrieval

We now formalize the problem of retrieving a desired object using a natural language query. Rather than constraining the description to a closed set of categories, a free-form text query $\mathbf{q} \in \mathcal{Q}$ is provided. For example, the user can search for *"the tennis ball"* or *"the Dove soap"* in Fig. 3.

The task is to identify which of a set of candidate images (or image regions) $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$ is the best match for the query $\mathbf{q}$. We assume each $\mathbf{c}_i$ contains a single object. We are therefore searching for a map $f_{\mathcal{C}} : \mathcal{Q} \to \{1, \ldots, k\}$. In particular, we compute a score $r(\mathbf{q}, \mathbf{c}_i)$ for each of the candidate objects and choose the one with the highest value:

$$f_{\mathcal{C}}(\mathbf{q}) = \underset{1 \le i \le k}{\operatorname{argmax}}\, r(\mathbf{q}, \mathbf{c}_i) \ .$$

In the case of a finite label space $\mathcal{Q}$, a standard vision baseline would regard the elements of $\mathcal{Q}$ as disjoint classes and learn classifiers for each of them. We could then choose the candidate object with the highest classifier score. However, in our scenario, we have unconstrained natural language queries, where learning a classifier for each element using traditional supervised learning is not an option, because not all query words could be observed at training time.

Therefore, our score function is based on comparing the given text query $\mathbf{q}$ with m different representations of an image in a weighted open-vocabulary text space. In particular, we define a set of functions $\Phi = \{\phi_j\}$, $j = 1, ..., m$, that project a given image $\mathbf{c}_i$ into a sparse vector of words, *i.e.* $S = \{(\mathbf{w}_n, \beta_n) | n \in \mathbb{N}\}$ with words $\mathbf{w}_n$ being the key and corresponding weights $\beta_n \in \mathbb{R}$ being the values of the sparse vector. We define a similar projection ψ for the given query. Each of the proposed projections ϕ_j results in a sparse representation based on the particular semantics that that specific function can extract from the image. In this paper we define five image-to-text projections, $\Phi = \{\phi_{IQE}, \phi_{GIS}, \phi_{DEC}, \phi_{LLC}, \phi_{CAF}\}$, where the first two are instance-level and the last three are category-level. Fig. 3 illustrates the respective strengths of category- and instance-based projections for several example objects.

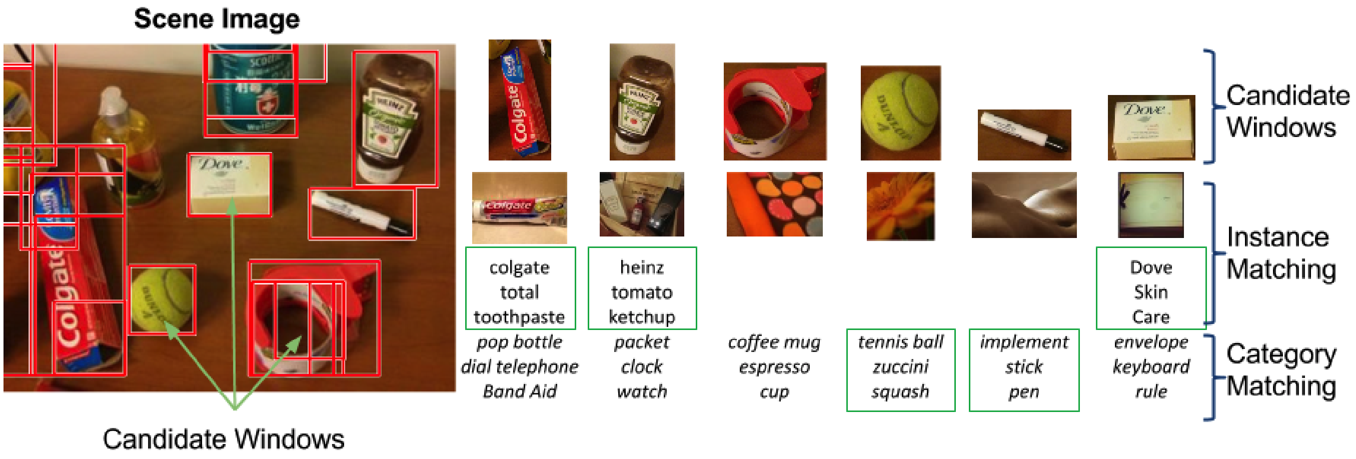


Fig. 3. Instance and category information are often complementary. The image on the left is overlaid with candidate image windows, computed using the selective search method of [53]. Extracted regions are shown on the top row. Below each region is the image that was the best match found by a web-scale instance search. Text below this matching image shows the corresponding instance-level projection; text below that (in italics) shows the category-level projection derived from Deep Convolutional Imagenet Classifiers (DECAF) (ϕ_{DEC}). In this example the instance-level projection would likely be able to resolve 3 of 6 objects for typical user queries in this scene (ketchup, toothpaste, bar of soap), while the category-level projection could likely resolve 2 of 6 (ball, pen). The liquid soap container was missed by the selective search in this example but is reasonably likely to have been recognized as a bottle by the DECAF models.

Once the images are projected into the weighted text space, we compute the similarity of each projection's weight vector $\phi_j(\mathbf{c}_i)$ to the query's weight vector $\psi(\mathbf{q})$. The similarities are combined across all projections to produce the final ranking using a cascade Cas:

$$r(\mathbf{q}, \mathbf{c}_i) = Cas(s(\psi(\mathbf{q}), \phi_1(\mathbf{c}_i)), ..., s(\psi(\mathbf{q}), \phi_m(\mathbf{c}_i)))$$

where $s(\cdot, \cdot)$ is the normalized correlation (or cosine angle). We describe each step of the algorithm below in detail.

We stress that our method is general and can accommodate other projections, such as projections that capture attribute-level semantics. For example, a variety of attribute projections could be defined, including those based on color [54], basic shapes, or based on surface markings such as text. For instance, one could incorporate OCR-based projections, as they provide a text attribute that is highly precise when it matches.

A. Category-based projections

We learned three category-based projections, ϕ_{LLC}, ϕ_{DEC}, ϕ_{CAF}, each with a different set of categories. The first one, ϕ_{LLC}, uses a bank of linear SVM classifiers over pooled local vector-quantized features learned from the 7,000 bottom level synsets of the 10K ImageNet database [15]. The second model, ϕ_{DEC}, makes use of the Deep Convolutional Network (DCN) developed and learned by [28] (the winning entry of the ILSVRC-2012 challenge) using the DECAF implementation [17]. The output layer consisted of 1,000 one-vs-all logistic classifiers, one for each of the ILSVRC-2012 object categories. The third model, ϕ_{CAF}, is a **new DCN for visual recognition** based on extending the DCN 1K ILSVRC-2012 of [17, 28] to a larger DCN, by replacing the last layer with 7,000 labels and then fine-tuning on the entire 7K Imagenet-2011 dataset. It was implemented with the Caffe framework (improved version of DECAF) available at http://caffe.berkeleyvision.org/.

We refer the interested reader to the corresponding publications for further details about these methods. We want to remark that Caffe and DECAF are open source, and that we are releasing the learned DCN models used in this work at http://openvoc.berkeleyvision.org, which are ready to be used by researchers working on robotics applications or on object retrieval.

Given the classification result, a traditional category-based approach would project an image to a vector with non-zero elements corresponding only to the text representation of its predicted label, e.g. *can*. However, only using a single label is likely to be error prone given the difficulty of category-based recognition. An image is therefore projected to the set of words $\mathbf{w}_n$ consisting of all synset synonyms, e.g. *can, tin, tin can*, with weights β_n corresponding to the corresponding predicted category probability. When the query description only consists of a single word, the resulting similarity score reduces to the sum of the predicted probabilities for the corresponding synsets.

More specifically, we define the LLC-10K projection as $\phi_{LLC}(\mathbf{c}_i) = \{(\mathbf{w}_n, p(\mathbf{w}_n|\mathbf{c}_i))\}$ with $\mathbf{w}_n$ being a word in a synset's list of synonyms and $p(\mathbf{w}_n|\mathbf{c}_i)$ being the posterior probability of the synsets where the word appear. A word can appear in more than one synset, so more frequent words would have a higher weight. To obtain the posterior probabilities for all the 10K synsets, we learn conventional one-vs-all classifiers on the leaf nodes, 7K in this case, obtain probability estimates for the leaves (via Platt scaling [44]), and then sum them to get the 3K internal node probabilities, as proposed in [16].

The DECAF-1K projection ϕ_{DEC} is defined similarly with the only difference being that the posterior probabilities for the 1K nodes are given directly by the output-layer of the deep architecture [17].

The CAFFE-7K projection ϕ_{CAF} is defined similarly with

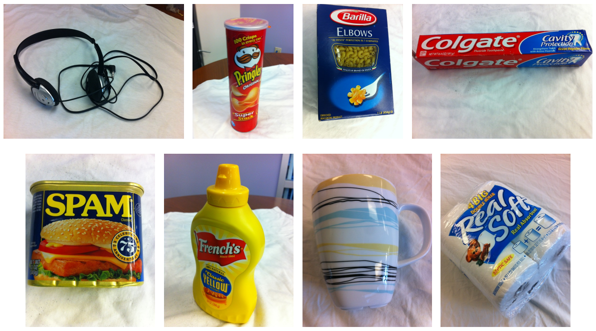


Fig. 4. Examples images from the *LAB* dataset.

the only difference being that the posterior probabilities for the leaves (7K) are given directly by the output-layer of the new learned DCN. All category projections ϕ_{LLC}, ϕ_{DEC}, ϕ_{CAF} project an image into a weighted set of 18K words, corresponding to all the words from the synset synonyms in 10K synsets used from WordNet.

B. Instance-based projection

The instance-based projections ϕ_{IQE} and ϕ_{GIS} used in our approach rely on large-scale image matching databases and algorithms which have been previously reported in the literature and have been available as commercial services for some time.

For ϕ_{IQE}, we use IQ Engines' (IQE) fully automated image matching API [3], which takes an image as input and provides a text output as a result, which is directly used as an image-to-text projection. The IQ Engines API indexes over one million images, mostly scraped from shopping webpages, using a local feature indexing with geometric verification paradigm [34, 40]. Each image in the database and each given query input image is represented by local features extracted at interest points. The first step of the matching is then to determine a candidate set by performing a k-nearest neighbor search using a visual bag-of-words signature computed from the local features. After obtaining the candidates in the product database, local feature matching is performed together with geometric validation and the description of the best matching image is returned. This technique can be seen as a version of the query expansion strategy of [12]. Given the best matched image, the corresponding product description is returned.

The ϕ_{GIS} projection is similar but based on the results of image-based queries to the Google Image Search (GIS) service. This service tries to match a given image with similar web images and returns a set of links in a fashion similar to the IQ Engines API service.

Both projections ϕ_{IQE} and ϕ_{GIS} are defined using a bag of words over the text returned by either IQ Engines or from the webpage summaries returned by Google Image Search. For example, for the image of the spam in Fig. 4, IQ Engines returns the following text ”Hormel Spam, Spam Oven Roasted Turkey”, while Google Image Search returns the best guess “spam” and links containing text like “do you use email in your business the can spam act establishes . . . ”.

C. Textual query expansion

The final projection ψ performs textual query expansion [10, 36] to relate brand names to corresponding object categories and also to tackle rare synonyms not present as synsets in ImageNet. Our textual query expansion technique is based on the large semantic concept database Freebase [9]. A given description $\mathbf{q}$ is parsed for noun groups using the standard NLP tagger and parser provided by the `nltk` framework[1]. A noun group could be, for example, the brand name ”cap’n crunch”. For each noun group, we query the Freebase database and substitute the non-synset noun group $\mathbf{w}$ with $\psi(\mathbf{w})$, if the query did return a result. The function ψ transforms $\mathbf{w}$ into a different set of words by searching for `/common/topic/description` entries in the Freebase results and concatenating them. After expansion, we can compare the projected weight vector with the weight vector obtained with one of the image-to-text projections ϕ_j described in the previous section. Note that more frequent words would have higher weight, as before. For example ”tazo chai tea” is expanded to ”Iced tea is a form of cold tea, usually served in a *glass* with ice . . . popular packaged *drink* . . . ” Here the italicized words are terms that are also found in the corresponding synset descriptions of the object to which the user was referring, thus this projection expands the query to include category-level words.

D. Cascade for combining similarities

We combine the similarities $s(\psi(\mathbf{q}), \phi_j(\mathbf{c}_i))$ computed for a candidate image $\mathbf{c}_i$ with a simple set of sequential decisions, using an optimized cascade Cas. Our cascade strategy works as follows: we sequentially process through our $j = 1, ..., m$ image-to-text projections, and if the jth similarity is informative, that is, when similarity for all $\mathbf{c}_i$ is not the same (within a small threshold), the result is returned, otherwise we continue with the next image-to-text projection. The order of the cascade is optimized using a greedy strategy, where the order of the similarity functions and the corresponding projection methods is determined by the Precision@1-NR (see Section IV-C). In our case, the first projection is based on IQ Engine, ϕ_{IQE}, which only outputs text in cases where a matching with a product image was successful. Zero scores of the instance-based similarity calculation typically occur when no matches are found by IQ-Engine or Google Image Search. The category-based methods result in zero similarity scores for examples where no category terms, *i.e.* words matching synsets in ImageNet, are part of the given query.

IV. Experiments

A. New open-vocabulary retrieval dataset

To quantitatively evaluate the proposed approaches, we collected natural language descriptions of images of objects in our laboratory (“Lab”) as well as from categories in the kitchen/household subtree of the ImageNet hierarchy (“Kitchen”). Fig. 4 and Fig. 3 illustrate the Lab images,

[1] http://nltk.org/

Method	P@1-NR	Coverage	P@1-All
MQ-Max [5]	60.62%	52.73%	40.34%
MQ-Avg [5]	58.57%	52.73%	38.86%
IQ-Engine (IQE)	**80.44%**	25.30%	32.41%
Google-Image (GIS)	69.88%	51.77%	44.22%
DECAF-1k (DEC)	67.70%	66.86%	50.93%
DEC+FB (DEC+)	61.71%	78.24%	**52.06%**
CAFFE-7k (CAF)	59.86%	79.94%	51.03%
LLC-10k (LLC)	57.89%	79.94%	49.80%
CAF+FB (CAF+)	54.14%	**88.66%**	50.04%
LLC+FB (LLC+)	52.63%	**88.66%**	48.63%

TABLE I
COMPARISON OF PROJECTIONS ON THE VALIDATION SET OF THE KITCHEN DATASET. P@1-NR: PRECISION@1 FOR NOT-RANDOM ANSWERS; COVERAGE: PERCENTAGE OF COVERED QUERIES; P@1-ALL: PRECISION@1 FOR ALL QUERIES

	Lab	Kitchen
Avg. number of words per description	3.34	4.70
Avg. number of nouns per description	2.19	2.73
Avg. number of adjectives per description	0.32	0.52
Avg. number of prepositions per description	0.27	0.50

TABLE II
STATISTICS OF THE DESCRIPTIONS WE OBTAINED FOR THE TWO DATASETS: DESCRIPTIONS WERE TAGGED WITH THE STANDARD PART-OF-SPEECH TAGGER IN NLTK

Method	Lab	Kitchen	ImageNet
MQ-Max [5]	35.26%	40.34%	43.43%
MQ-Avg [5]	33.40%	38.86%	41.42%
IQ-Engine (IQE)	48.59%	32.85%	24.46%
Google-Image (GIS)	48.30%	44.45%	39.19%
DECAF-1K (DEC)	43.36%	50.73%	53.13%
DEC+FB (DEC+)	42.19%	52.13%	54.05%
CAFFE-7K (CAF)	44.70%	51.34%	57.50%
CAF+FB (CAF+)	42.19%	50.04%	56.82%
LLC-10K (LLC)	40.05%	49.57%	57.24%
LLC+FB (LLC+)	37.85%	41.25%	56.27%
Linear-SVM (LSVM)	45.75%	58.90%	63.65%
Rank-SVM (RSVM)	56.40%	62.49%	72.62%
Max-Kernel (MAX)	49.51%	61.11%	68.49%
IQE,GIS	56.37%	51.86%	58.86%
IQE,GIS,DEC	64.09%	60.95%	75.04%
IQE,GIS,DEC,CAF	66.45%	64.15%	80.13%
IQE,GIS,DEC,CAF,LLC	66.76%	65.10%	81.50%
Full Cascade (CAS)	**67.07%**	**66.20%**	**81.93%**

TABLE III
PRECISION@1-ALL THE QUERIES FOR THE THREE EXPERIMENTS AND FOR ALL THE METHODS.

Method	Category	Instance	Unlabeled
MQ-Max [5]	27.65%	51.61%	36.52%
MQ-Avg [5]	26.09%	49.18%	31.96%
IQ-Engine (IQE)	40.41%	67.18%	52.93%
Google-Image (GIS)	40.46%	66.69%	53.95%
DECAF-1K (DEC)	48.15%	25.42%	48.38%
CAFFE-7K (CAF)	48.80%	30.79%	44.20%
LLC-10K (LLC)	44.00%	25.92%	42.94%
Linear-SVM (LSVM)	47.67%	39.13%	43.17%
Rank-SVM (RSVM)	54.33%	69.73%	58.76%
Max-Kernel (MAX)	50.08%	49.10%	52.63%
Full Cascade (CAS)	**62.92%**	**76.58%**	**65.95%**

TABLE IV
DETAILED ANALYSIS OF PRECISION@1 FOR THE LAB EXPERIMENT BY TYPE OF QUERY. AMONG THE 1830 QUERIES, 53% WERE LABELED AS CATEGORY, 18% WERE LABELED AS INSTANCE AND THE REST REMAINED UNLABELED

while Fig. 1 illustrates the Kitchen set. Each image was posted on Amazon Mechanical Turk in order to collect natural language descriptions. For each image, ten individuals were asked to provide a free-form description of the object in the image as though they were instructing a robot to go through the house and locate it, *e.g.*"Robot, please bring me the ** fill in the blank **". The descriptions we obtained are fairly rich and diverse and TABLE II contains some statistics. There are 183 images annotated in the Lab set and 606 images annotated in the Kitchen set, additionally there are 74240 images that serve as distractors. Given that for each annotated image there are 10 annotations, for our evaluations we used over 60K combinations of targets, annotations and distractors.

To support the detailed evaluation below, each query provided was additionally labeled by a second annotator as to whether it appeared to be an "instance" or a "category"-level query. These were selected on the basis of the textual description without looking at the image they were given for. The category- and instance-level labels were applied when a query had a brand-name or fine-grained description or had a clear category term directly related to the synset, respectively. The other queries remained unlabeled. The dataset will be made publicly available.

We created a series of synthetic trials to simulate the scenario shown in Fig. 3. We sample a query image from the Lab or Kitchen sets and a number of distractors from the same set or from all of ImageNet ("ImageNet"), the latter being a considerably easier task. The descriptions associated with the query image serve as the object retrieval query. For each pair of target image and textual description, we sample 10 image distractors from different synsets, obtaining 6060 trials comprised of 11 images (one is the target) and one text description.

B. Baselines

We evaluate the three categorical methods DECAF-1K (DEC), CAFFE-7K (CAF), and LLC-10K (LLC) from Section III-A, the two instance-based methods IQ-Engine (IQE) and Google-Image-Search (GIS) from Section III-B, and their combinations using Linear-SVM (LSVM), Rank-SVM (RSVM), Max-Kernel (MAX) and Cascade (Cas) as given in Section III-D. The inputs to the SVMs are the scores from each of the projections, and they are trained to choose the target object over the distractors. For instance to train the Linear-SVM we labeled the targets as positive and the distractors as negatives, while to train the Rank-SVM we imposed the constraints that the targets should be ranked above all the distractors. The output of the trained models is a global score computed as a weighted combination of the individual projections. Furthermore, we also show that our Freebase query expansion proposed in Section III-C helps to improve the overall accuracy.

We also compare all methods with the best Multi-Query approach of [5], where a given query description is given to Google image search (not to be confused with the search-by-image service GIS we are using) and the similarity of the images with the candidate images is estimated with a visual bag-of-words pipeline. We refer to the resulting methods as Multiple Queries Max (MQ-Max) and Multiple Queries Average (MQ-Avg) depending on the pooling performed.

C. Experimental setup

To analyze the methods in detail, we have defined the following performance measures: (1) *Coverage* is the percentage of trials in which the method given the text query and the images is able to give an informative answer, that is, the cases in which it produces different values for the candidates, and therefore the target selection is not random. (2) *Precision@1-NR (Not-Random)* is the precision of the 1-st ranked image, computed only on the trials described in (1), i.e. where the method is able to deterministically select the target object. (3) *Precision@1-All* measures Precision@1 for all cases including cases where the method guesses the target randomly since it cannot determine which one is the target.

To learn the parameters of the combined methods Linear-SVM (LSVM), Rank-SVM (RSVM) we used a small validation set comprised of 100 target images with their corresponding textual descriptions and distractors. To establish the order for the sequential Cascade method, we order the individual methods by their Precision@1-NR on the validation set from the highest to the lowest (TABLE I).

D. Comparing individual projection methods

First, we analyze each image-to-text projection in isolation. The results are given in TABLE I for the Kitchen dataset, *i.e.* kitchen domain images from ImageNet used for the target as well as distractor images.

The method with the highest Precision@1-NR value but lowest coverage is IQ-Engine (IQE), which means that the method is very precise when a match is found but also likely not to return anything (zero similarity values to the candidate images). The methods with the highest coverage are LLC-10k (LLC) which has the lowest precision, and CAFFE-7K (CAF) which has a slightly higher precision, meaning that these method are likely able to allow for proper candidate selection, but are not as precise as IQE.

We can also see that the Freebase query expansion technique we proposed in Section III-C mainly increases the coverage but reduces the precision, which is an intuitive result because the number of keywords in a query increases significantly due to expansion.

E. Combining image-to-text projections

The main results of our object retrieval experiments are given in TABLE III for our lab images and the kitchen domain images from ImageNet with distractor images from the same domain or random ones sampled from other synsets of ImageNet not necessarily related to the kitchen domain.

The best individual method depends on the experiment; for instance in the Lab experiment, the best is IQ-Engine (IQE) with P@1-All 48.59%, in the Kitchen experiment, the best one is DECAF-1K+Freebase (DEC+) with P@1-All 52.13%, and in the ImageNet experiment, the best one is CAFFE-7K (CAF+) with P@1-All 57.50%.

However, the best combined method is consistently the Full Cascade (CAS), with P@1-All 67.07% for Lab, with P@1-All 66.20% for Kitchen and with P@1-All 81.93% for ImageNet. The Full Cascade includes the Freebase query expansion,and obtains a substantial performance gain in all the experiments.

The best individual method varies across all three datasets, but we are able to outperform the method of [5] in all cases. More importantly, we are able to combine all similarity and projection methods with our cascade combination, which outperforms all individual methods and other combinations. In particular the last rows of TABLE III show how each method when added to to the sequence it improves the performance.

F. Which method is helping for which type of query?

As can be seen in Fig. 3, and in the results of the combined methods show in TABLE III, the category and instance-level methods benefit from each other in the combination and can be thus considered as orthogonal concepts. A further proof for this fact can be seen in detail when looking on the results for queries labeled as category or instance-level queries in the dataset, which are given in TABLE IV.

The instance-level projections IQE and GIS show a higher Precision@1-All on the instance queries than the category-based projections DECAF-1K, CAFFE-7K and LLC-10K and vice versa for the category queries. Among the category-based projections, CAFFE-7K has the highest Precision@1-All on the category and instance queries, showing the better capabilities of the new trained DCN to handle fine-grained object recognition.

G. Runtime discussion

For robotics applications, where runtime is an important issue when making predictions, we suggest to use CAFFE-7K as a category-based projection, since it offers very fast prediction with around 2s per 256 test images, including 1.5s to read and preprocess them (using 4 cores) and 0.5s to run the DCN in Caffe (using a Titan GPU, 6s in CPU mode). The additional runtime for GIS, IQE and the Freebase query expansion depends on the speed of the proprietary web service, but was in the order of a few seconds in our experiments. Furthermore, our approach offers easy parallelization by distributing the different modules on several machines. Taking into account that in total we are dealing with a recognition system learned with several million images and thousands of categories, this is a remarkable runtime and a great opportunity for improving robotics applications that need to deal with everyday-life objects.

V. Conclusions and Future Work

We have proposed an architecture for open-vocabulary object retrieval based on image-to-text projections from components across varying semantic levels. We have shown empirically that a combined approach which fuses category-level and instance-level projections outperforms existing baselines and either projection alone on user queries which refer to one of a number of objects of interest.

Key aspects of our method include that: 1) images are matched not simply to a pre-defined class label space but retrieved using a multi-word descriptive phrase; 2) query expansion for unusual terms improves performance; 3) instance matching can improve category-level retrieval and vice-versa.

Our framework is general and can be expanded to include other projections defined on attributes based on color, text cues, and other modalities that are salient for a domain. In our opinion, our approach is extremely useful for robotics applications, because we are the first ones to combine several of the most powerful visual recognition techniques available today: deep neural networks trained on ImageNet and large-scale image matching. Our framework is just the beginning of an open source project in open-vocabulary object retrieval and we will provide source code and pre-trained models ready to use for robotics applications at http://openvoc.berkeleyvision.org.

References

[1] Freebase API. https://developers.google.com/freebase/.

[2] Google image search. http://www.images.google.com/.

[3] IQ Engines: Image Recognition APIs for photo albums and mobile commerce. https://www.iqengines.com/.

[4] R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *WWW*, 2013.

[5] Relja Arandjelovic and Andrew Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012.

[6] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *ICCV*, 2001.

[7] A. Berg, R. Farrell, A. Khosla, J. Krause, L. Fei-Fei, L. Jia, and S. Maji. Fine-Grained Challenge 2013. https://sites.google.com/site/fgcomp2013/, 2013.

[8] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR*, 2003.

[9] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, 2008.

[10] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *CSUR*, 2012.

[11] Ken Chatfield and Andrew Zisserman. Visor: Towards on-the-fly large-scale object category retrieval. In *ACCV 2012*, pages 432–446. Springer, 2013.

[12] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007.

[13] Jeffrey Dalton, James Allan, and Pranav Mirajkar. Zero-shot video retrieval using content and concepts. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1857–1860. ACM, 2013.

[14] Thomas Dean, Mark A Ruzon, Mark Segal, Jonathon Shlens, Sudheendra Vijayanarasimhan, and Jay Yagnik. Fast, accurate detection of 100,000 object classes on a single machine. In *CVPR*, 2013.

[15] Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What Does Classifying More Than 10,000 Image Categories Tell Us? In *ECCV*, 2010.

[16] Jia Deng, Jonathan Krause, Alex Berg, and Li Fei-Fei. Hedging Your Bets: Optimizing Accuracy-Specificity Trade-offs in Large Scale Visual Recognition. In *CVPR*, 2012.

[17] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *ArXiv e-prints*, 2013.

[18] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*. Springer, 2010.

[19] Ryan Farrell, Om Oza, Ning Zhang, Vlad I Morariu, Trevor Darrell, and Larry S Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, 2011.

[20] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *NIPS*, 2013.

[21] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[22] Albert Gordoa, José A Rodríguez-Serrano, Florent Perronnin, and Ernest Valveny. Leveraging category-level labels for instance-level image retrieval. In *CVPR*, 2012.

[23] David Grangier and Samy Bengio. A discriminative kernel-based model to rank images from text queries. *PAMI*, 2007.

[24] S. Guadarrama, L. Riano, D. Golland, D. Gohring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell. Grounding spatial relations for human-robot interaction. In *IROS*, 2013.

[25] S. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *TPAMI*, 2012.

[26] J. Krapac, M. Allan, J. Verbeek, and F. Jurie. Improving

web image search results using query-relative classifiers. In *CVPR*, 2010.

[27] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1(2):193–206, 2013.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[29] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011.

[30] Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. Generalizing image captions for image-text parallel corpus. In *ACL*, 2013.

[31] Xirong Li, Dong Wang, Jianmin Li, and Bo Zhang. Video search in concept subspace: a text-like paradigm. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 603–610. ACM, 2007.

[32] Yuanqing Lin, Fengjun Lv, Shenghuo Zhu, Kai Yu, Ming Yang, and Timothee Cour. Large-scale image classification: fast feature extraction and SVM training. In *CVPR*, 2011.

[33] Yiming Liu, Dong Xu, and Ivor W. Tsang. Using large-scale web data to facilitate textual query based retrieval of consumer photos. In *ACM-MM*, 2009.

[34] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[35] Aurelien Lucchi and Jason Weston. Joint image and word sense discrimination for image retrieval. In *ECCV*, 2012.

[36] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[37] Pierre Moreels, Michael Maire, and Pietro Perona. Recognition by probabilistic hypotheis construction. In *ECCV*, 2004.

[38] Apostol Paul Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the 15th international conference on Multimedia*, pages 991–1000. ACM, 2007.

[39] Shi-Yong Neo, Jin Zhao, Min-Yen Kan, and Tat-Seng Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *Image and Video Retrieval*, pages 143–152. Springer, 2006.

[40] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[41] Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. From large scale image categorization to entry-level categories. In *ICCV*, 2013.

[42] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Cats and dogs. In *CVPR*, 2012.

[43] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[44] John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large-margin Classifiers*, 1999.

[45] Nikhil Rasiwasia, Pedro J Moreno, and Nuno Vasconcelos. Bridging the gap: Query by semantic example. *Multimedia, IEEE Transactions on*, 9(5):923–938, 2007.

[46] Abhishek Sharma, Abhishek Kumar, H Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012.

[47] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[48] Cees GM Snoek, Bouke Huurnink, Laura Hollink, Maarten De Rijke, Guus Schreiber, and Marcel Worring. Adding semantics to detectors for video retrieval. *Multimedia, IEEE Transactions on*, 9(5):975–986, 2007.

[49] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *EMNLP*, 2012.

[50] Jie Tang, Stephen Miller, Arjun Singh, and Pieter Abbeel. A textured object recognition pipeline for color and depth image data. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3467–3474. IEEE, 2012.

[51] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.

[52] Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Toward a probabilistic approach to acquiring information from human partners using language. Technical report, MIT, 2012.

[53] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV*, 2013.

[54] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. Learning color names for real-world applications. *TIP*, 2009.

[55] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

[56] Ziang Xie, Arjun Singh, Justin Uang, Karthik S Narayan, and Pieter Abbeel. Multimodal blending for high-accuracy instance recognition. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 2214–2221. IEEE, 2013.