The International Journal of Robotics Research
000(00):1–20
©The Author(s) 2010
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI:doi number
http://mms.sagepub.com

Understanding Object Descriptions in Robotics by Open-vocabulary Object Retrieval and Detection

Sergio Guadarrama*[1], Erik Rodner[2], Kate Saenko[3] and Trevor Darrell[1]
[1]*EECS Department, University of California at Berkeley, USA*
[2]*Computer Vision Group, Friedrich Schiller University of Jena, Germany*
[3]*CS Department, University of Massachussetts Lowell, USA*

Abstract

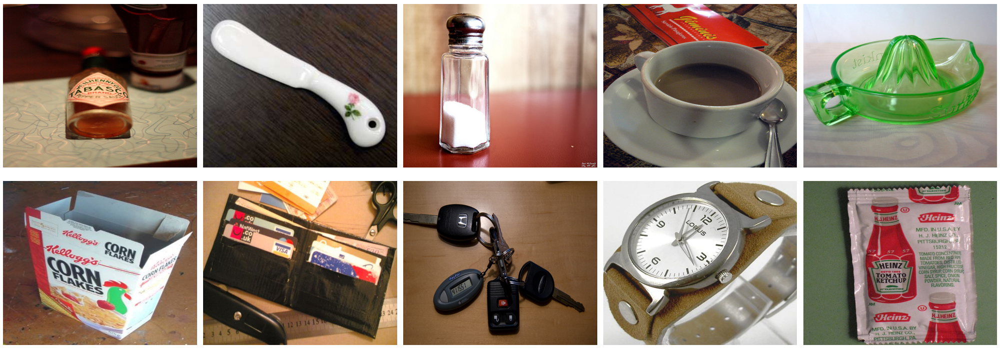
We address the problem of retrieving and detecting objects based on open-vocabulary natural language queries: Given a phrase describing a specific object, e.g., "the corn flakes box", the task is to find the best match in a set of images containing candidate objects. When naming objects, humans tend 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. Furthermore, we show how to process queries referring to objects within scenes, using state-of-the-art adapted detectors. 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.

Keywords
convolutional neural networks, natural language processing, object retrieval, object detection

1. 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

[*] Corresponding author; e-mail: sguada@eecs.berkeley.edu



“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.

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 (Furnas et al. 1987).

Large-vocabulary object recognition has recently made significant advances, spurred by the emergence of dictionary-scale datasets (Deng et al. 2010, Lin et al. 2011). Dramatic progress has been made on category-level recognition (Krizhevsky et al. 2012), 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 subcategories, e.g. *sparrow, Prius, Coke bottle* (Berg et al. 2013). Addressing the *open-vocabulary* problem that arises when natural language strings are the label space requires recognizing potentially millions of separate categories (Agrawal et al. 2013). In the robotic perception setting, detecting object instances often requires good RGB-D models of the actual objects to be recognized (Tang et al. 2012, Xie et al. 2013), and therefore do not scale well to more than a hundred objects.

This paper is an extended and reviewed version of Guadarrama et al. (2014). In this paper following on our previous work 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, 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.

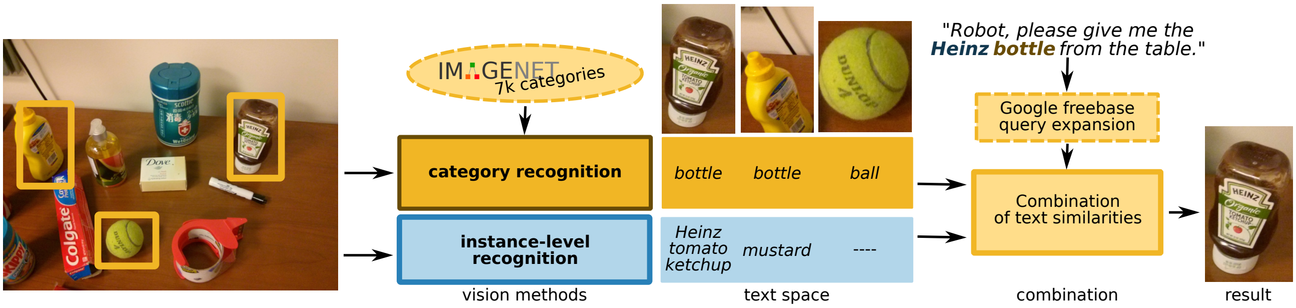

IMAGENET
7k categories
category recognition
instance-level recognition
vision methods
bottle
bottle
ball
Heinz tomato ketchup
mustard
----
text space
"Robot, please give me the Heinz bottle from the table."
Google freebase query expansion
Combination of text similarities
combination
result

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.

While *generating* descriptive nouns, attributes, and/or phrases from an image is a topic of recent interest (Kuznetsova et al. 2013, Deng et al. 2012, Ordonez et al. 2013), the challenge of *retrieving* an image or object from within a scene using natural language has had less attention (but see Guadarrama et al. (2013), Tellex et al. (2012, 2011)). 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 (Weston et al. 2011, Socher et al. 2012, Frome et al. 2013). 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 (Deng et al. 2010) and a deep convolutional model trained on ImageNet (Krizhevsky et al. 2012), including a version that can recognize 7,400 different objects, 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 (IQEngines 2014, Google 2014) 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 (Freebase 2014), 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. In our experiments, we presume that object bounding boxes are known (in a deployed system see Section 5, we rely on bottom-up segmentation or "objectness" to provide a region shortlist).

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 Arandjelovic & Zisserman (2012).

2. 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 (Deng et al. 2010), and output a single category label (Deng et al. 2012, Krizhevsky et al. 2012), while others try to generate proper natural language descriptions (Farhadi et al. 2010). Recent efforts have investigated increasingly fine-grained recognition approaches (Farrell et al. 2011, Parkhi et al. 2012), 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 (Lowe 1999, Moreels et al. 2004, Philbin et al. 2007, Sivic & Zisserman 2003), 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 (Tang et al. 2012, Xie et al. 2013). 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 (Krishnamurthy & Kollar 2013) 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 (Dean et al. 2013), 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 (Tang et al. 2012, Xie et al. 2013), in that it handles novel objects and out-of-vocabulary labels.

Earlier work has focused on modeling co-occurrences between image regions and tags (Barnard & Forsyth 2001, Blei & Jordan 2003), focusing on scenes where the correspondence between image regions and tags is unknown. Hwang and Grauman (Hwang & Grauman 2012) 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 (Sharma et al. 2012) 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. (2011), *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 (Grangier & Bengio 2007, Krapac et al. 2010, Liu et al. 2009, Lucchi & Weston 2012). 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, Gordoa et al. (2012) 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.

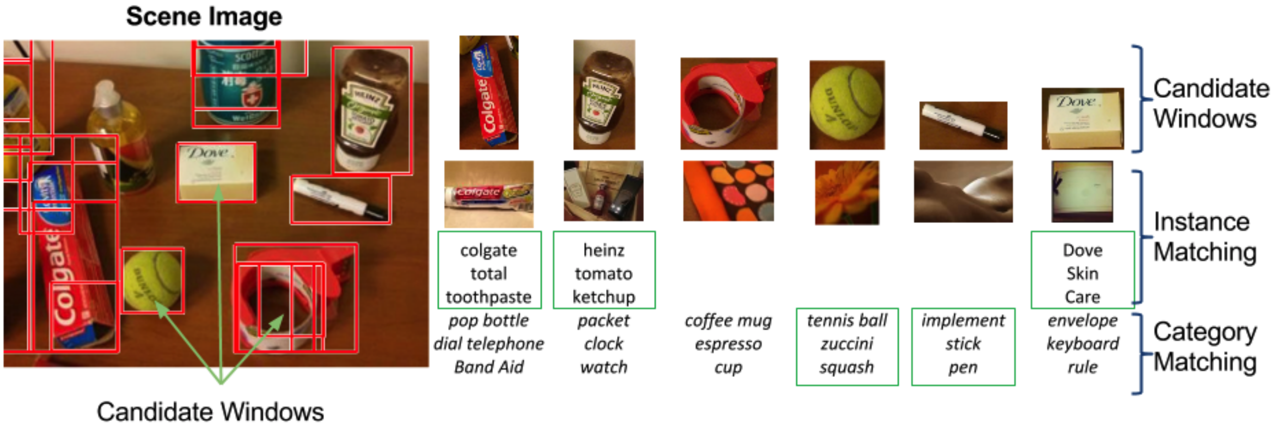


Fig. 3. Instance and category information are often complementary. The image on the left is overlaid with candidate image windows, computed using the selective search method of Uijlings et al. (2013). Extracted regions are shown on the top row. Below each region is the image that was the best match found 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.

Another line of work within image retrival is (Arandjelovic & Zisserman 2012, Chatfield & Zisserman 2013), where authors try to find a set of relevant images in a dataset given a short text query. Arandjelovic & Zisserman (2012) 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. Chatfield & Zisserman (2013) 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 (Li et al. 2007, Natsev et al. 2007, Rasiwasia et al. 2007, Snoek et al. 2007, Neo et al. 2006); more recently (Dalton et al. 2013) addressed the problem of zero-shot video retrieval using content and concepts.

3. 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:

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

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.

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 (Van De Weijer et al. 2009), 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.

3.1. 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 (Deng et al. 2010). The second model, ϕ_{DEC}, makes use of the Deep Convolutional Network (DCN) developed and learned by Krizhevsky et al. (2012) (the winning entry of the ILSVRC-2012 challenge) using the DECAF implementation (Donahue et al. 2013). 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 Donahue et al. (2013) and Krizhevsky et al. (2012) to a larger DCN, by replacing the last layer with 7,400 labels and then fine-tuning on the 7K Imagenet-2009 fall release. It was implemented with the Caffe framework (improved version of DECAF) available at `http://caffe.berkeleyvision.org/`.

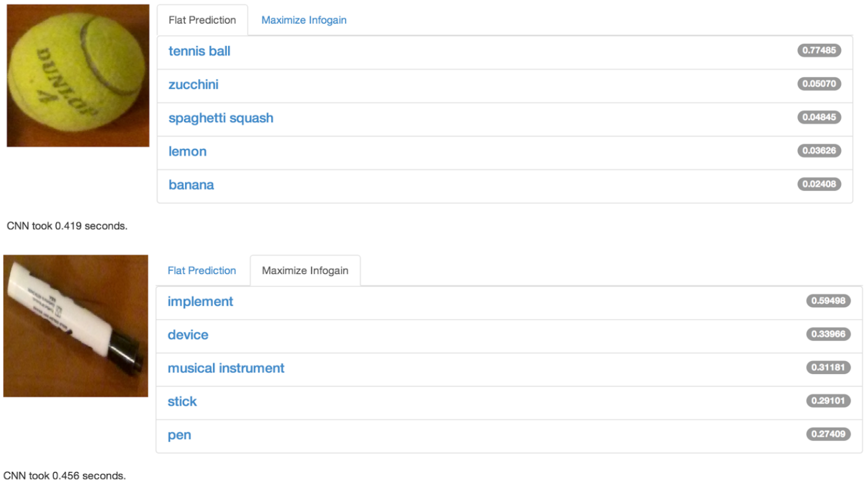
We refer the interested reader to the corresponding publications for further details about these methods. The methods we use are indeed partially redundant and within a robotics application one would restrict them to just a few category-based restrictions. However, in this paper, we also aim at comparing different projections and their suitability for open-vocabulary object retrieval (Section 4.4). 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 Platt (1999)), and then sum them to get the 3K internal node probabilities, as proposed in Deng et al. (2012).

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 (Donahue et al. 2013).

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



Flat Prediction | Maximize Infogain

tennis ball	0.77485
zucchini	0.05070
spaghetti squash	0.04845
lemon	0.03626
banana	0.02408

CNN took 0.419 seconds.

Flat Prediction | Maximize Infogain

implement	0.59498
device	0.33966
musical instrument	0.31181
stick	0.29101
pen	0.27409

CNN took 0.456 seconds.

Fig. 4. Examples of category recognition with state-of-the-art CNN-based approaches: `http://caffe.berkeleyvision.org`.

3.2. *Instance-based projection*

While category-level recognition has received a lot of attention in the last decade, it was rarely combined with ideas from instance-level or exemplar-based approaches outside image retrieval. This can be attributed to the high intra-class variations of the categories commonly used to evaluate the methods, variations that make restrictive exemplar-based matching problematic. However, especially indoors, we are surrounded by all kinds of products, where most of the intra-class variation originates from different viewpoints and capture conditions. This fact can be exploited by exemplar-based approaches efficiently and allows for exact matching of objects parts and geometric reasoning.

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.

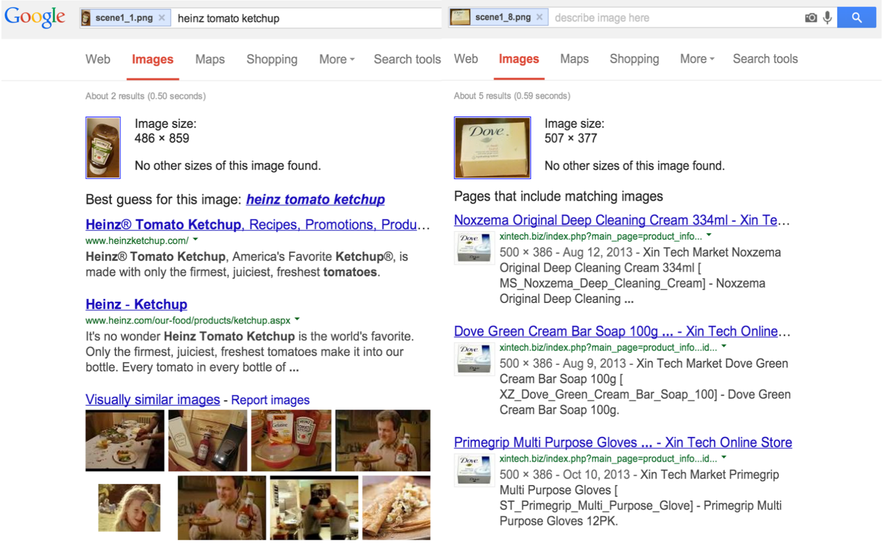
For ϕ_{IQE}, we use IQ Engines' (IQE) fully automated image matching API IQEngines (2014)[1], which takes an image as input and provides a text output as a result, which is directly used as an image-to-text projection. The IQ Engines API indexes over one million images, mostly scraped from shopping webpages, using local feature indexing with a geometric

[1] IQ Engines has since been aqcuired by Yahoo Inc.; similar services include Google's search-by-image, which we also evaluate in this study, and CamFind (http://camfindapp.com)

verification paradigm Lowe (1999), Nister & Stewenius (2006). 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 Chum et al. (2007). 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. 7, 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 . . . ". Additional examples are shown in Fig. 5.



About 2 results (0.50 seconds)

Image size:
486 × 859

No other sizes of this image found.

Best guess for this image: ***heinz tomato ketchup***

Heinz® Tomato Ketchup, Recipes, Promotions, Produ…
www.heinzketchup.com/
Heinz® Tomato Ketchup, America's Favorite **Ketchup®**, is made with only the firmest, juiciest, freshest **tomatoes**.

Heinz - Ketchup
www.heinz.com/our-food/products/ketchup.aspx
It's no wonder **Heinz Tomato Ketchup** is the world's favorite. Only the firmest, juiciest, freshest tomatoes make it into our bottle. Every tomato in every bottle of **...**

Visually similar images - Report images

About 5 results (0.59 seconds)

Image size:
507 × 377

No other sizes of this image found.

Pages that include matching images

Noxzema Original Deep Cleaning Cream 334ml - Xin Te…
xintech.biz/index.php?main_page=product_info...
500 × 386 - Aug 12, 2013 - Xin Tech Market Noxzema Original Deep Cleaning Cream 334ml [MS_Noxzema_Deep_Cleaning_Cream] - Noxzema Original Deep Cleaning ...

Dove Green Cream Bar Soap 100g ... - Xin Tech Online…
xintech.biz/index.php?main_page=product_info...id...
500 × 386 - Aug 9, 2013 - Xin Tech Market Dove Green Cream Bar Soap 100g [XZ_Dove_Green_Cream_Bar_Soap_100] - Dove Green Cream Bar Soap 100g.

Primegrip Multi Purpose Gloves ... - Xin Tech Online Store
xintech.biz/index.php?main_page=product_info...id...
500 × 386 - Oct 10, 2013 - Xin Tech Market Primegrip Multi Purpose Gloves [ST_Primegrip_Multi_Purpose_Glove] - Primegrip Multi Purpose Gloves 12PK.

Fig. 5. Examples of instance matching with Google Image Search's search-by-image service.

3.3. *Textual query expansion*

The final projection ψ performs textual query expansion Carpineto & Romano (2012), Manning et al. (2008) 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 Bollacker et al. (2008). A given description $\mathbf{q}$ is parsed for noun groups using the standard NLP tagger and parser provided by the `nltk` framework[2]. 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

[2] http://nltk.org/

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.

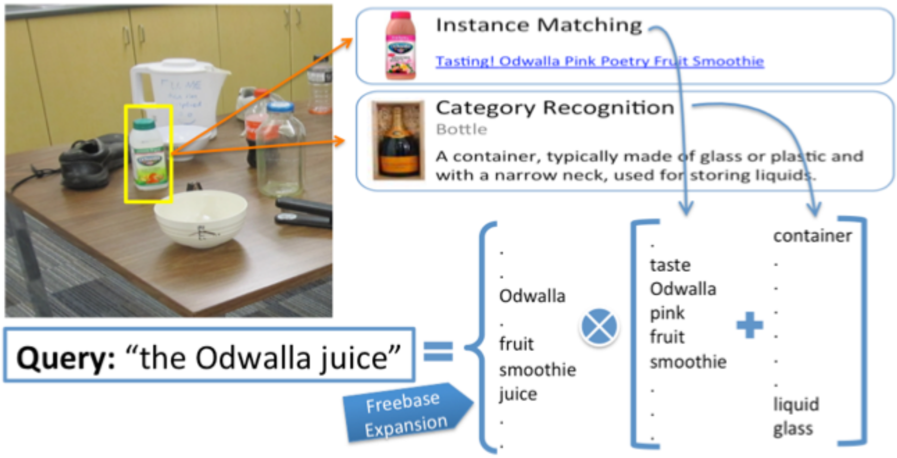

Instance Matching
Tasting! Odwalla Pink Poetry Fruit Smoothie
Category Recognition
Bottle
A container, typically made of glass or plastic and with a narrow neck, used for storing liquids.
Query: "the Odwalla juice"
Freebase Expansion
Odwalla
fruit
smoothie
juice
taste
Odwalla
pink
fruit
smoothie
container
liquid
glass

Fig. 6. Overview of approach: to evaluate how closely an image region matches a user's text query, the query is first expanded using Freebase to include words like "fruit" and "smoothie", then compared with the combined text vectors obtained by the instance and category projections of the image region.

3.4. Combining similarities: Max-Kernel, Linear-SVM, Rank-SVM and Cascade

As illustrated in Fig. 6, the text vectors $\phi_j(\mathbf{c}_i)$ obtained by the various image-to-text projections of the image $\mathbf{c}_i$ can be combined (e.g., summed) and then used to compute similarity with the expaned query vector $\psi(\mathbf{q})$. In our experiments we combine the similarities $s(\psi(\mathbf{q}), \phi_j(\mathbf{c}_i))$ for each projection j computed for each candidate image $\mathbf{c}_i$ using different methods. Then the aggregated similarity is used to rank the candidate regions and pick the most similar to the query.

Max-Kernel (MAX) compute the overall similarity using the maximum across all the individual projections (we also tried other functions like average or minimum, but they performed worse). We also used a Linear-SVM (LSVM) and a Rank-SVM (RSVM) to learn how to combine the individual similarities. 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 (see Joachims (2002) for details). The output of the trained models is a global score computed as a weighted combination of the individual projections.

However, based on the observation that different methods have very different precision/recall behavior in this section we propose a more efficient and accurate approach. We combine the individual similarities 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 4.3). 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.

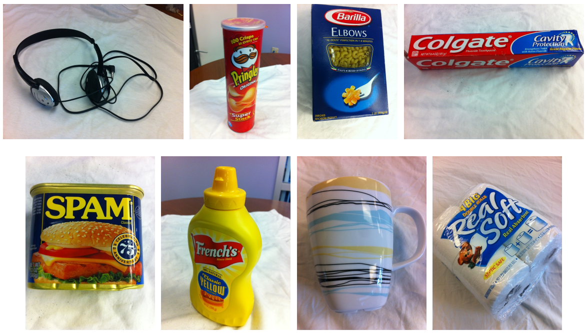


Fig. 7. Example images from the *LAB* dataset.

	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 1. Statistics of the descriptions we obtained for the two datasets[3]

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.

4. Experiments

4.1. 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. 7 and Fig. 3 illustrate the Lab images, 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 1 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 instance- and category-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

[3] Descriptions were tagged using the standard part-of-speech tagger in `nltk`

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.

4.2. Baselines

We evaluate the three categorical methods DECAF-1K (DEC), CAFFE-7K (CAF), and LLC-10K (LLC) from Section 3.1, the two instance-based methods IQ-Engine (IQE) and Google-Image-Search (GIS) from Section 3.2, and their combinations using Linear-SVM (LSVM), Rank-SVM (RSVM) and Max-Kernel (MAX) as given in Section 3.4. Furthermore, we also show that our Freebase (FB) query expansion proposed in Section 3.3 helps to improve the overall accuracy (denoted with '+' sign in the table).

We also compare all the proposed methods with the best Multi-Query approach of Arandjelovic & Zisserman (2012), where a given query description is given to Google image search and the similarity of the images with the candidate images is estimated with a visual bag-of-words pipeline (using the code made available by Vedaldi (2014) and the parameters specified in the paper: for each query we use the top 8 results given by Google Image Search and encode them using 1M visual words, for our experiments we used spatial re-ranking for all the comparisons). We refer to the resulting methods as Multiple Queries Max (MQ-Max) and Multiple Queries Average (MQ-Avg) depending on the pooling performed. This baseline should not to be confused with the search-by-image service GIS we are using for instance matching, where we upload each image and get a textual description back (when there is a match) and then compare the results with the query description.

4.3. 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 2).

4.4. Comparing individual projection methods

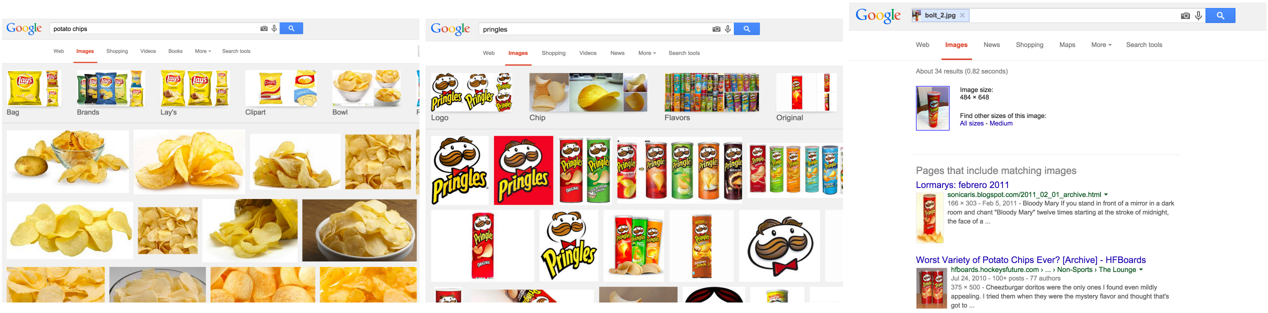
First, we analyze each image-to-text projection in isolation. The results are given in Table 2 for the Kitchen dataset, *i.e.* kitchen domain images from ImageNet used for the target as well as distractor images.

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.

Based on the results of Table 2 the instance-based projections (IQE and GIS) have higher Precision@1-NR (they tend to be correct when they provide an answer), while category-based projections (DEC-1k, CAF-7k and LLC-10k) have higher recall (especially when using Freebase query expansion). We can also see that the Freebase query expansion

Method	P@1-NR	Coverage	P@1-All
MQ-Max (Arandjelovic & Zisserman 2012)	60.62%	52.73%	40.34%
MQ-Avg (Arandjelovic & Zisserman 2012)	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 2. 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



About 34 results (0.82 seconds)

Image size:
484 × 648

Find other sizes of this image:
All sizes - Medium

Pages that include matching images

Lormarys: febrero 2011
sonicaris.blogspot.com/2011_02_01_archive.html
166 × 303 - Feb 5, 2011 - Bloody Mary If you stand in front of a mirror in a dark room and chant "Bloody Mary" twelve times starting at the stroke of midnight, the face of a ...

Worst Variety of Potato Chips Ever? [Archive] - HFBoards
hfboards.hockeysfuture.com › ... › Non-Sports › The Lounge
Jul 24, 2010 - 100+ posts - 77 authors
375 × 500 - Cheezburgar doritos were the only ones I found even mildly appealing. I tried them when they were the mystery flavor and thought that's got to ...

Fig. 8. Images retrieved by Google Image Search search-by-text for the queries *"potato chips"* and *"pringles"* (first and second from the left) vs the results retrieved via search-by-image given the input image of a Pringles can (far right).

technique we proposed in Section 3.3 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.

Although both GIS and MQ-Max/Avg rely on Google Image Search, the first one uses search-by-image while the second one uses search-by-text functionality. To illustrate the difference in their behavior let us consider the results of two different queries, "potato chips" or "pringles", using search-by-text (as used by MQ-Max and MQ-Avg). They return two very different sets of images (see Fig. 8) that will make the algorithm to fail to recognize the image of the Pringles when the query is "potato chips". On the contrary, given the image of the Pringles, GIS with search-by-image will find a match for Pringles and potato chips and handle both queries correctly. Another important difference between our instance matching methods (IQ and GIS) and the multi-query methods (MQ-Max or MQ-Avg) is the size of the pool of candidate images, while the instance matching relies on millions images, multi-query methods relies on a few downloaded images. As can be seen in Table 2 GIS has better precision than MQ-Max and MQ-Avg, while having similar coverage.

According to Table 3 the instance-based projections (IQE and GIS) perform better in the Lab and Kitchen experiments, where there are more products and instances. They perform worse in the ImageNet experiment, where there are less products and more generic objects. On the other hand, the category-based projections (DEC-1k, CAF-7k and LLC-10k) perform better on the ImageNet experiment, with CAF-7k and LLC-10k performing best due to broad coverage. However, they perform worse in the Lab experiment.

Method	Lab	Kitchen	ImageNet
MQ-Max (Arandjelovic & Zisserman 2012)	35.26%	40.34%	43.43%
MQ-Avg (Arandjelovic & Zisserman 2012)	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 3. Precision@1-All the queries for the three experiments and for all the methods.

4.5. Combining image-to-text projections

The main results of our object retrieval experiments are given in Table 3 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 table itself also gives a good summary of the results of individual methods, although only focusing on the P@1-All performance measure.

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, we are able to outperform the method of Arandjelovic & Zisserman (2012) 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. 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. Other more sophisticated combination techniques lead to an inferior performance. In the case of LSVM and Rank-SVM this can be contributed to a weak linear combination model that tries to combine very heterogenous scores.

The last rows of Table 3 show how each method when added to the sequence improves the performance. Since the components are added by decreasing order of precision, the following model will only be responsible for the cases where the previous model could not provide an answer. It can be seen that instance-based projections IQ and GIS are quite complementary and combining them provide a significant improve across all the experiments (see IQE,GIS).

Although all the category-based projections (DEC-1k, CAF-7k and LLC-10k) try to capture a good representation of categories, they use different features (see Section 3.1 for details) and an increasing number of categories from 1k up to 10k. These differences allow for diversity in the combination and improve performance. Furthermore, the queries people used to refer to objects often either contain a category or an instance-type description. All of the individual methods alone (without freebase) can handle only a single type of queries. Combining their strengths is not only reasonable but also straightforward with our simple cascade technique. The biggest jump of the performance can be observed when the first

Method	Category	Instance	Unlabeled
MQ-Max (Arandjelovic & Zisserman 2012)	27.65%	51.61%	36.52%
MQ-Avg (Arandjelovic & Zisserman 2012)	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 4. 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.

category-based method (DEC) is added to the cascade (Table 3), the performances improves over 8 percent points for all datasets. After this we only get slight improvements from different category-based methods.

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

To further analyze the difference in performance between methods we conducted a detailed analysis of Precision@1 for the Lab experiment by type of query (see Table 4). From this table it can be seen that MQ-Max and MQ-Avg perform much better when the query contains an Instance (i.e "Pringles") than when it contains a Category (i.e. "potato chips") describing the object of interest (see Fig. 8). Also we can see the importance of using a large scale database of images to get good instance matching, and the results obtained by IQE and GIS are significally better that MQ-Max and MQ-Avg even for queries containing Instances.

Overall 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.

As can be seen in Fig. 3, and in the results of the combined methods show in Table 3, 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 4.

4.7. 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.

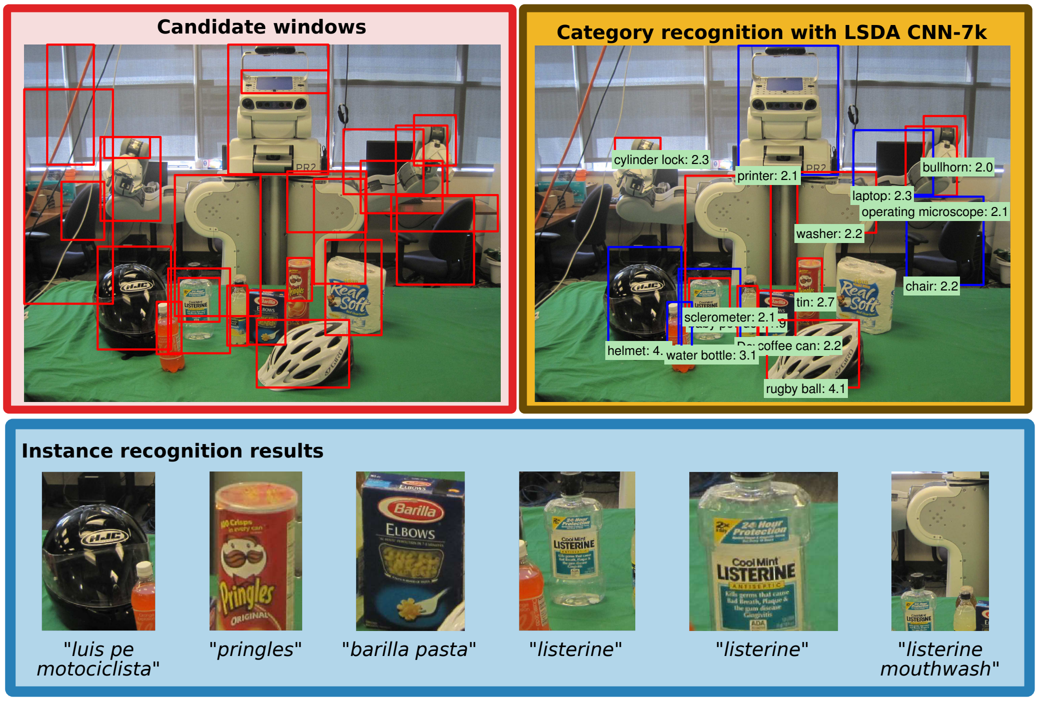


Fig. 9. *Robot scene*: Detection results of the category and instance recognition methods.

5. Open-vocabulary Object Detection

In previous sections and in our previous work Guadarrama et al. (2014), we had assumed that our algorithm operates in classification mode, i.e. it receives cropped images as input, each containing a single object. In practical scenarios, we would want it to also locate objects in larger scenes, i.e. to operate in detection mode. In the following, we show results obtained by combining state-of-the-art detection methods with our open-vocabulary techniques. This leads to an end-to-end open-vocabulary detection pipeline, which provides an extremely powerful tool for robotics.

Both classification and detection are key visual recognition challenges, though historically very different architectures have been deployed for each. Very recently, object detection methods have made significant advances in terms of performance (Girshick et al. 2014) as well as in terms of the number of categories that can be detected simultaneously (Hoffman et al. 2014).

5.1. Open-vocabulary Object Retrieval with Adapted RCNN Detectors

The recent RCNN detection model (Girshick et al. 2014) has shown state-of-the-art performance on the Pascal VOC challenge (Everingham et al. 2010) and on the ImageNet200 detection challenge (Russakovsky et al. 2014). However, these challenges only contain 20 and 200 categories, respectively. Unlike large-scale classification training data which we used for our classifiers above, detection training data for other categories is generally unavailable. For our open vocabulary object retrieval, we need to be able to detect (localize and classify) thousands of different categories. Fortunately, the LSDA method of Hoffman et al. (2014) proposes an adaptation method to transform object classifiers into RCNN object detectors. Using their method, we are able to transform our CAFFE-7k object classifiers into a LSDA-7k object detectors. As before, we also ascend the hierachy from the leaf nodes to get predictions for all 10k categories. In the following, we briefly describe RCNN as well as the LSDA technique.

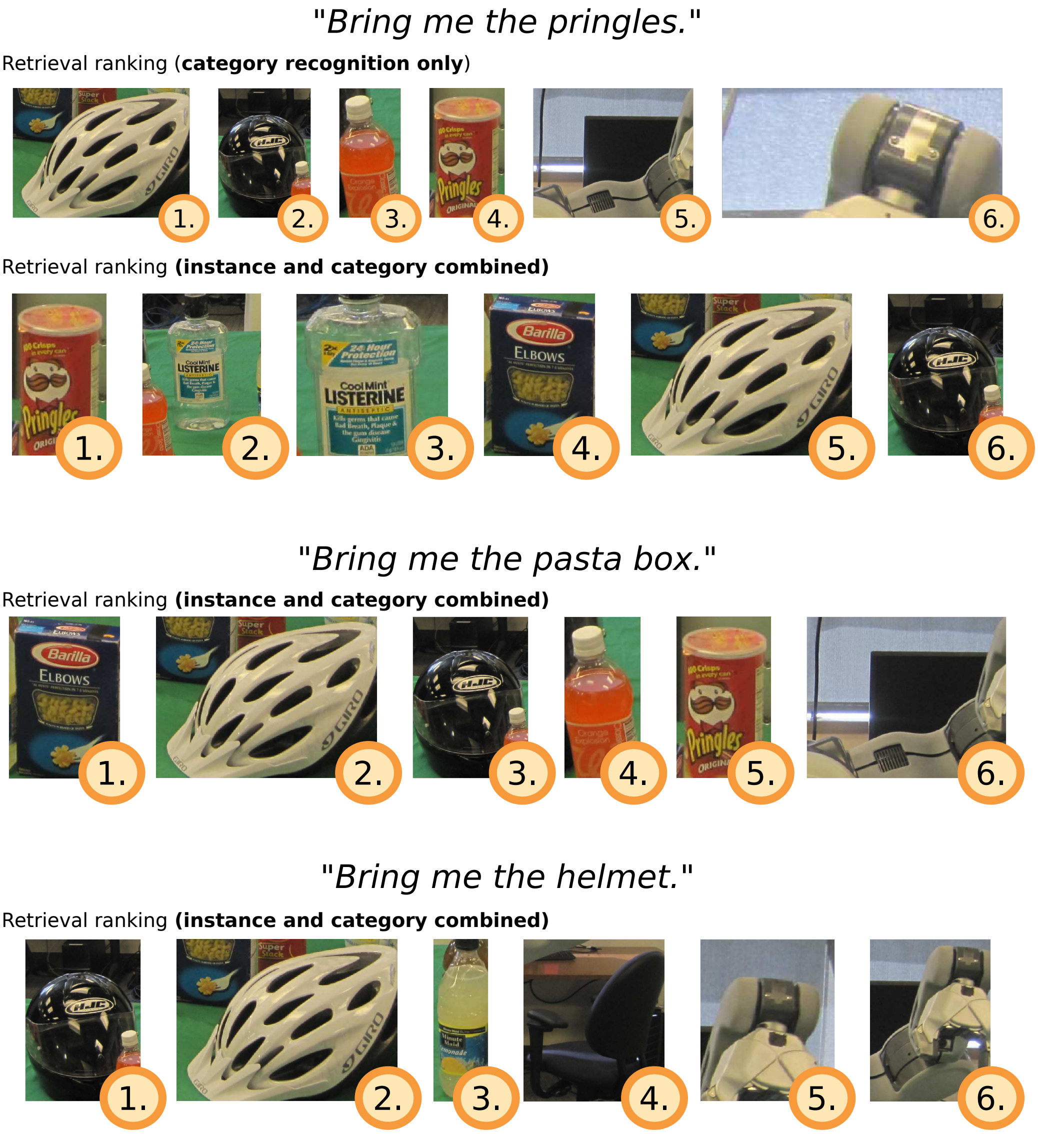

"Bring me the pringles."
Retrieval ranking (category recognition only)
1.
2.
3.
4.
5.
6.
Retrieval ranking (instance and category combined)
1.
2.
3.
4.
5.
6.
"Bring me the pasta box."
Retrieval ranking (instance and category combined)
1.
2.
3.
4.
5.
6.
"Bring me the helmet."
Retrieval ranking (instance and category combined)
1.
2.
3.
4.
5.
6.

Fig. 10. Retrieval results for the *robot scene* (Fig. 9) and three different user queries.

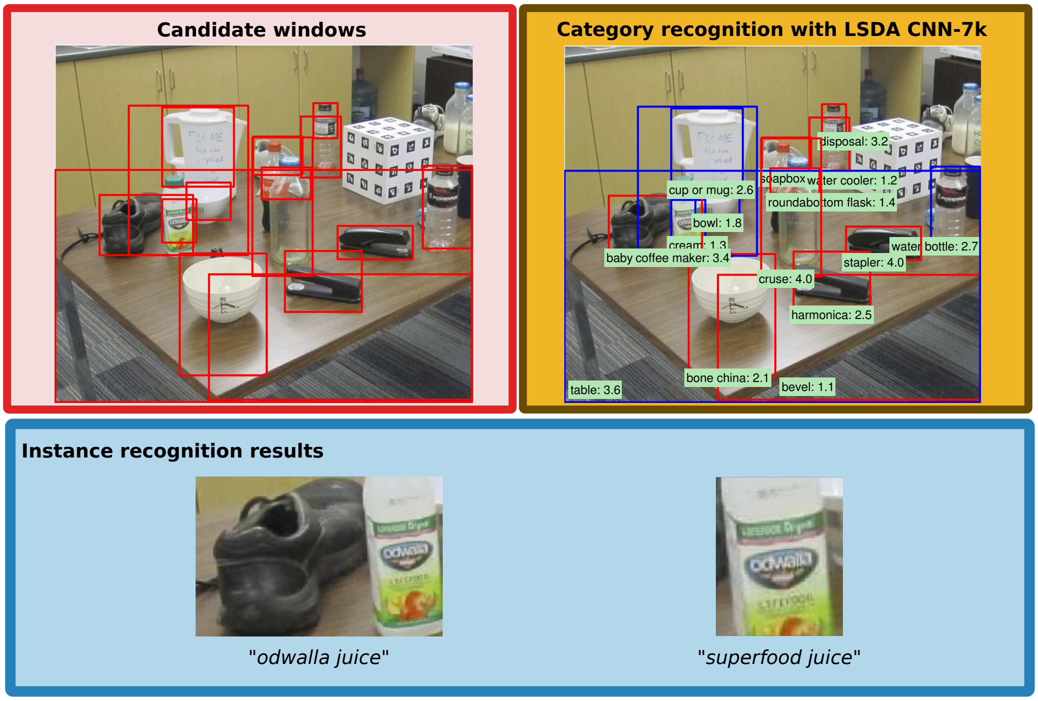


Fig. 11. *Table scene*: Detection results of the category and instance recognition methods.

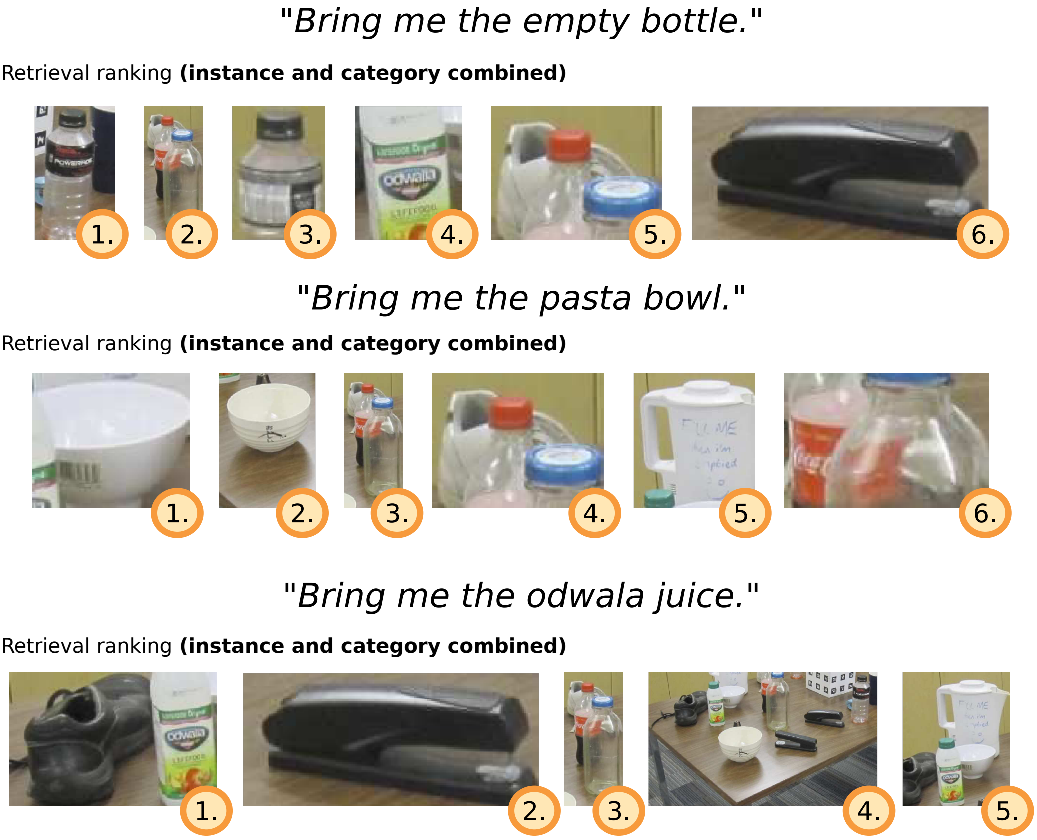

"Bring me the empty bottle."
Retrieval ranking (instance and category combined)
1.
2.
3.
4.
5.
6.
"Bring me the pasta bowl."
Retrieval ranking (instance and category combined)
1.
2.
3.
4.
5.
6.
"Bring me the odwala juice."
Retrieval ranking (instance and category combined)
1.
2.
3.
4.
5.

Fig. 12. Retrieval results for the *table scene* (Fig. 11) and three different user queries.

Region Proposals for RCNN The RCNN detector of Girshick et al. (2014) obtains automatic object proposals (bounding boxes likely to contain objects) from a scene, and then classifies them with a convolutional neural network. Object proposal generation has been an active area of research in computer vision in recent years (Arbeláez et al. 2014, Krähenbühl & Koltun 2014, Zitnick & Dollár 2014, Uijlings et al. 2013). Given the importance of good region proposals Hosang et al. (2014), researchers have studied the problem of using depth images to improve the quality of object proposals (Lin et al. 2013, Gupta et al. 2014). Gupta et al. (2014), used depth information to obtain improved contours from RGB-D images via a multiscale combinatorial grouping framework (Arbeláez et al. 2014) to report great improvements over RGB only methods, obtaining the same recall with an order of magnitude fewer regions as compared to RGB only methods.

In the experimental results in Section 5.2, we used the selective search method of Uijlings et al. (2013), which performs well with a running time of 2 seconds per image. Faster alternatives include the geodesic object proposal approach of Krähenbühl & Koltun (2014), since it can obtain good proposals in under 0.25 seconds. Furthermore, a method that uses depth information like Gupta et al. (2014) could be used if RGB-D data were available to improve performance and if a processing time of around 12 seconds per image is not a critical roadblock for the application.

Large-Scale Detection through Adaptation (LSDA) It is much cheaper and easier to collect large quantities of classification training data with image-level labels from search engines than it is to collect detection training data and label it with precise bounding boxes. The Large-Scale Detection through Adaptation (LSDA) algorithm learns the difference between the two tasks and transfers this knowledge to classifiers for categories without bounding box annotated data, turning them into detectors. This can be thought of as adaptation of object models from a classifier domain to a detection domain. First, LSDA learns convolutional neural network classifiers (AlexNet) for all categories in the 7k ImageNet classification training database. Next, it learns detectors for categories in the labeled detection training database by finetuning all layers of the CNN classification network and adding a background class. Finally, for those of the 7k categories without detection data (all but 200) it adapts the classification output layer (last layer of the neural network) using a transformation based the change in the output layer for the known detection categories. The result is a convolutional neural network that can take region proposals and classify them as either on of the 7k categories, or background. The final step is non-maximum suppression across all detected objects in the scene. We refer the reader to Hoffman et al. (2014) for further details of the algorithm.

5.2. *Results and evaluation*

We integrated the LSDA-7k results and ran our instance recognition methods on each object proposal, to obtain an end-to-end pipeline for open vocabulary object retrieval. Due to the lack of space, we only show the results on two very complex scenes, which we refer to as *robot* (Fig. 9) and *table* scene (Fig. 11).

The pure results of the category and the instance recognition are shown in Fig. 9 for the *robot* scene. The LSDA-7k category recognition results are displayed in the right image along with their respective bounding boxes and they are surprisingly accurate for some categories. However, they are not able to provide very fine-grained and detailed annotations for example in the case of the “pringles box”. In contrast, the instance recognition approach only gives results for six object proposals but with instance-level captions, such as “pringles”. The final object retrieval results can be seen in Fig. 10 as rankings of the object proposals. Whereas the category-only object retrieval fails for the task “Bring me the pringles”, the combination with instance recognition allows for obtaining the right object. For the “pasta box” task a similar conclusion can be drawn. In the case of the “helmet”, the LSDA-7k result already provides a perfect match with the query.

Fig. 11 and Fig. 12 show another complex scenario, where we are able to handle very different and challenging queries.

6. Conclusions

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`.

7. Future Work

While our approach takes significant steps towards allowing natural interaction between a user and a robotic agent, in the following, we describe several future research directions to further improve its performance and to integrate it into realistic robotics systems.

Cues from Gesture Recognition Gestures can play an important role in reducing the number of ambiguities during object retrieval. Let us assume that a robust gesture recognition system is available, such as the ones presented in Pateraki et al. (2014). One idea to integrate gesture cues would be then to first compute a probability distribution from a pointing gesture to estimate where the user likely pointed to. This estimation does not need to have a high precision (low entropy distribution), since we could simply use the values as weights for each hypothesis. Combining these weights with the text similarities we are currently computing, would allow for selecting the object that fits to the user query and the pointing gesture.

Novelty Detection and Active Learning Our algorithm currently always assign one of the given hypotheses to the query. What if the hypothesis detection is not correct or there is simply no object (on the table) that matches the query? An important research direction will be to extend our algorithm to allowing for rejections of given queries. Whereas this could be achieved by setting a minimum threshold for the text similarities, it would be beneficial to incorporate ideas from the area of novelty detection Bodesheim et al. (2013). Furthermore, active learning and classification techniques could allow for an improved human-machine interaction Kding et al. (2015), Freytag et al. (2014).

Pose Estimation and Grasping Finding the object location is a useful capability, however, further work is needed to integrate it into a real robotics system that can manipulate objects. In the future, we would like to incorporate our approach as a first step in an object manipulation system that processes natural language input from a user. Such as system would need to perform tasks like fine-grained pose estimation and grasp planning. For example, the user could interact with the system to correct its mistakes, or instruct it to use certain grasping techniques for certain objects, see for example Ralph & Moussa (2008). We also envision extending our approach to use 3D point cloud data to match objects to extensive online CAD model databases.

Acknowledgements

This work was supported in part by DARPA's MSEE and SMISC programs, NSF awards IIS-1427425, IIS-1212798, and IIS-1116411, Toyota, DAAD, and the Berkeley Vision and Learning Center.

References

Agrawal, R., Gupta, A., Prabhu, Y. & Varma, M. (2013), Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages, *in* 'International Conference on World Wide Web (WWW)', pp. 13–24.

Arandjelovic, R. & Zisserman, A. (2012), Multiple queries for large scale specific object retrieval., *in* 'BMVC'.

Arbeláez, P., Pont-Tuset, J., Barron, J. T., Marques, F. & Malik, J. (2014), Multiscale combinatorial grouping, *in* 'Computer Vision and Pattern Recognition (CVPR)'.

Barnard, K. & Forsyth, D. (2001), Learning the semantics of words and pictures, *in* 'International Conference on Computer Vision (ICCV)'.

Berg, A., Farrell, R., Khosla, A., Krause, J., Fei-Fei, L., Jia, L. & Maji, S. (2013), 'Fine-Grained Challenge 2013', https://sites.google.com/site/fgcomp2013/.

Blei, D. M. & Jordan, M. I. (2003), Modeling annotated data, *in* 'International ACM SIGIR conference on Research and development in informaion retrieval', pp. 127–134.

Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M. & Denzler, J. (2013), Kernel null space methods for novelty detection, *in* 'IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', pp. 3374–3381.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T. & Taylor, J. (2008), Freebase: a collaboratively created graph database for structuring human knowledge, *in* 'SIGMOD'.

Carpineto, C. & Romano, G. (2012), 'A survey of automatic query expansion in information retrieval', *ACM Computing Surveys (CSUR)* **44**(1), 1:1–1:50.

Chatfield, K. & Zisserman, A. (2013), Visor: Towards on-the-fly large-scale object category retrieval, *in* 'ACCV 2012', Springer, pp. 432–446.

Chum, O., Philbin, J., Sivic, J., Isard, M. & Zisserman, A. (2007), Total recall: Automatic query expansion with a generative feature model for object retrieval, *in* 'International Conference on Computer Vision (ICCV)'.

Dalton, J., Allan, J. & Mirajkar, P. (2013), Zero-shot video retrieval using content and concepts, *in* 'Proceedings of the 22nd ACM international conference on Conference on information & knowledge management', ACM, pp. 1857–1860.

Dean, T., Ruzon, M. A., Segal, M., Shlens, J., Vijayanarasimhan, S. & Yagnik, J. (2013), Fast, accurate detection of 100,000 object classes on a single machine, *in* 'Computer Vision and Pattern Recognition (CVPR)'.

Deng, J., Berg, A. C., Li, K. & Fei-Fei, L. (2010), What Does Classifying More Than 10,000 Image Categories Tell Us?, *in* 'ECCV'.

Deng, J., Krause, J., Berg, A. & Fei-Fei, L. (2012), Hedging Your Bets: Optimizing Accuracy-Specificity Trade-offs in Large Scale Visual Recognition, *in* 'Computer Vision and Pattern Recognition (CVPR)'.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. & Darrell, T. (2013), 'DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition', *ArXiv e-prints* .

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

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J. & Forsyth, D. (2010), Every picture tells a story: Generating sentences from images, *in* 'European Conference on Computer Vision (ECCV)', Springer.

Farrell, R., Oza, O., Zhang, N., Morariu, V. I., Darrell, T. & Davis, L. S. (2011), Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance, *in* 'International Conference on Computer Vision (ICCV)'.

Freebase (2014), 'Freebase API', https://developers.google.com/freebase/.

Freytag, A., Rodner, E. & Denzler, J. (2014), Selecting influential examples: Active learning with expected model output changes, *in*

‘European Conference on Computer Vision (ECCV)’, Vol. 8692, pp. 562–577.

Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M. & Mikolov, T. (2013), DeViSE: A deep visual-semantic embedding model, *in* ‘Advances in Neural Information Processing Systems (NIPS)’.

Furnas, G. W., Landauer, T. K., Gomez, L. M. & Dumais, S. T. (1987), ‘The vocabulary problem in human-system communication’, *Communications of the ACM* **30**(11), 964–971.

Girshick, R., Donahue, J., Darrell, T. & Malik, J. (2014), Rich feature hierarchies for accurate object detection and semantic segmentation, *in* ‘Computer Vision and Pattern Recognition (CVPR)’.

Google (2014), ‘Google image search.’, http://www.images.google.com/.

Gordoa, A., Rodríguez-Serrano, J. A., Perronnin, F. & Valveny, E. (2012), Leveraging category-level labels for instance-level image retrieval, *in* ‘Computer Vision and Pattern Recognition (CVPR)’.

Grangier, D. & Bengio, S. (2007), ‘A discriminative kernel-based model to rank images from text queries’, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* .

Guadarrama, S., Riano, L., Golland, D., Gohring, D., Jia, Y., Klein, D., Abbeel, P. & Darrell, T. (2013), Grounding spatial relations for human-robot interaction, *in* ‘IROS’.

Guadarrama, S., Rodner, E., Saenko, K., Zhang, N., Farrell, R., Donahue, J. & Darrell, T. (2014), Open-vocabulary object retrieval, *in* ‘Proceedings of Robotics: Science and Systems (RSS)’, Berkeley, USA.
URL: *http://www.eecs.berkeley.edu/ sguada/pdfs/2014-RSS-open-vocabulary-final.pdf*

Gupta, S., Girshick, R., Arbeláez, P. & Malik, J. (2014), Learning rich features from rgb-d images for object detection and segmentation, *in* ‘European Conference on Computer Vision (ECCV)’, Springer, pp. 345–360.

Hoffman, J., Guadarrama, S., Tzeng, E., Donahue, J., Girshick, R., Darrell, T. & Saenko, K. (2014), ‘LSDA: Large scale detection through adaptation’, arXiv:1407.5035.

Hosang, J., Benenson, R. & Schiele, B. (2014), How good are detection proposals, really?, *in* ‘British Machine Vision Conference (BMVC)’.

Hwang, S. & Grauman, K. (2012), ‘Reading between the lines: Object localization using implicit cues from image tags’, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* .

IQEngines (2014), ‘IQ Engines: Image Recognition APIs for photo albums and mobile commerce.’, https://www.iqengines.com/.

Joachims, T. (2002), Optimizing search engines using clickthrough data, *in* ‘Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining’, ACM, pp. 133–142.

Krähenbühl, P. & Koltun, V. (2014), Geodesic object proposals, *in* ‘European Conference on Computer Vision (ECCV)’, Springer, pp. 725–739.

Krapac, J., Allan, M., Verbeek, J. & Jurie, F. (2010), Improving web image search results using query-relative classifiers, *in* ‘Computer Vision and Pattern Recognition (CVPR)’.

Krishnamurthy, J. & Kollar, T. (2013), ‘Jointly learning to parse and perceive: Connecting natural language to the physical world’, *Transactions of the Association for Computational Linguistics* **1**(2), 193–206.

Krizhevsky, A., Sutskever, I. & Hinton, G. (2012), Imagenet classification with deep convolutional neural networks, *in* ‘Advances in Neural Information Processing Systems (NIPS)’.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C. & Berg, T. L. (2011), Baby talk: Understanding and generating simple image descriptions, *in* ‘Computer Vision and Pattern Recognition (CVPR)’.

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T. & Choi, Y. (2013), Generalizing image captions for image-text parallel corpus, *in* ‘ACL’.

Kding, C., Freytag, A., Rodner, E., Bodesheim, P. & Denzler, J. (2015), Active learning and discovery of object categories in the presence of unnameable instances, *in* ‘IEEE Conference on Computer Vision and Pattern Recognition (CVPR)’.

Li, X., Wang, D., Li, J. & Zhang, B. (2007), Video search in concept subspace: a text-like paradigm, *in* ‘Proceedings of the 6th ACM international conference on Image and video retrieval’, ACM, pp. 603–610.

Lin, D., Fidler, S. & Urtasun, R. (2013), Holistic scene understanding for 3D object detection with RGBD cameras, *in* 'International Conference on Computer Vision (ICCV)'.

Lin, Y., Lv, F., Zhu, S., Yu, K., Yang, M. & Cour, T. (2011), Large-scale image classification: fast feature extraction and SVM training, *in* 'Computer Vision and Pattern Recognition (CVPR)'.

Liu, Y., Xu, D. & Tsang, I. W. (2009), Using large-scale web data to facilitate textual query based retrieval of consumer photos, *in* 'ACM-MM'.

Lowe, D. G. (1999), Object recognition from local scale-invariant features, *in* 'International Conference on Computer Vision (ICCV)'.

Lucchi, A. & Weston, J. (2012), Joint image and word sense discrimination for image retrieval, *in* 'European Conference on Computer Vision (ECCV)'.

Manning, C. D., Raghavan, P. & Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA.

Moreels, P., Maire, M. & Perona, P. (2004), Recognition by probabilistic hypotheis construction, *in* 'European Conference on Computer Vision (ECCV)'.

Natsev, A. P., Haubold, A., Tešić, J., Xie, L. & Yan, R. (2007), Semantic concept-based query expansion and re-ranking for multimedia retrieval, *in* 'Proceedings of the 15th international conference on Multimedia', ACM, pp. 991–1000.

Neo, S.-Y., Zhao, J., Kan, M.-Y. & Chua, T.-S. (2006), Video retrieval using high level features: Exploiting query matching and confidence-based weighting, *in* 'Image and Video Retrieval', Springer, pp. 143–152.

Nister, D. & Stewenius, H. (2006), Scalable recognition with a vocabulary tree, *in* 'Computer Vision and Pattern Recognition (CVPR)'.

Ordonez, V., Deng, J., Choi, Y., Berg, A. C. & Berg, T. L. (2013), From large scale image categorization to entry-level categories, *in* 'ICCV'.

Parkhi, O. M., Vedaldi, A., Jawahar, C. V. & Zisserman, A. (2012), Cats and dogs, *in* 'Computer Vision and Pattern Recognition (CVPR)'.

Pateraki, M., Baltzakis, H. & Trahanias, P. (2014), 'Visual estimation of pointed targets for robot guidance via fusion of face pose and hand orientation', *Computer Vision and Image Understanding* **120**, 1–13.

Philbin, J., Chum, O., Isard, M., Sivic, J. & Zisserman, A. (2007), Object retrieval with large vocabularies and fast spatial matching, *in* 'Computer Vision and Pattern Recognition (CVPR)'.

Platt, J. C. (1999), Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, *in* 'Advances in Large-margin Classifiers'.

Ralph, M. & Moussa, M. (2008), 'Toward a natural language interface for transferring grasping skills to robots', *Robotics, IEEE Transactions on* **24**(2), 468–475.

Rasiwasia, N., Moreno, P. J. & Vasconcelos, N. (2007), 'Bridging the gap: Query by semantic example', *IEEE Transactions on Multimedia* **9**(5), 923–938.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. & Fei-Fei, L. (2014), 'ImageNet Large Scale Visual Recognition Challenge'.

Sharma, A., Kumar, A., Daume, H. & Jacobs, D. W. (2012), Generalized multiview analysis: A discriminative latent space, *in* 'Computer Vision and Pattern Recognition (CVPR)'.

Sivic, J. & Zisserman, A. (2003), Video Google: A text retrieval approach to object matching in videos, *in* 'International Conference on Computer Vision (ICCV)'.

Snoek, C. G., Huurnink, B., Hollink, L., De Rijke, M., Schreiber, G. & Worring, M. (2007), 'Adding semantics to detectors for video retrieval', *IEEE Transactions on Multimedia* **9**(5), 975–986.

Socher, R., Huval, B., Manning, C. D. & Ng, A. Y. (2012), Semantic compositionality through recursive matrix-vector spaces, *in* 'EMNLP'.

Tang, J., Miller, S., Singh, A. & Abbeel, P. (2012), A textured object recognition pipeline for color and depth image data, *in* 'Robotics and Automation (ICRA), 2012 IEEE International Conference on', IEEE, pp. 3467–3474.

Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S. & Roy, N. (2011), Understanding natural language commands for robotic navigation and mobile manipulation, *in* 'AAAI'.

Tellex, S., Thaker, P., Deits, R., Simeonov, D., Kollar, T. & Roy, N. (2012), Toward a probabilistic approach to acquiring information from human partners using language, Technical report, MIT.

Uijlings, J., van de Sande, K., Gevers, T. & Smeulders, A. (2013), 'Selective search for object recognition', *International Journal of Computer Vision (IJCV)* **104**(2), 154–171.

Van De Weijer, J., Schmid, C., Verbeek, J. & Larlus, D. (2009), 'Learning color names for real-world applications', *Transactions on Image Processing* .

Vedaldi, A. (2014), 'Visualindex - a simple image indexing engine in matlab', `https://github.com/vedaldi/visualindex`.

Weston, J., Bengio, S. & Usunier, N. (2011), Wsabie: Scaling up to large vocabulary image annotation, *in* 'IJCAI'.

Xie, Z., Singh, A., Uang, J., Narayan, K. S. & Abbeel, P. (2013), Multimodal blending for high-accuracy instance recognition, *in* 'Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on', IEEE, pp. 2214–2221.

Zitnick, C. L. & Dollár, P. (2014), Edge boxes: Locating object proposals from edges, *in* 'European Conference on Computer Vision (ECCV)', Springer, pp. 391–405.