Int J Comput Vis (2012) 100:134–153
DOI 10.1007/s11263-011-0494-3

Learning the Relative Importance of Objects from Tagged Images for Retrieval and Cross-Modal Search

Sung Ju Hwang · Kristen Grauman

Received: 16 December 2010 / Accepted: 23 August 2011 / Published online: 18 October 2011

Abstract We introduce an approach to image retrieval and auto-tagging that leverages the implicit information about object *importance* conveyed by the list of keyword tags a person supplies for an image. We propose an unsupervised learning procedure based on Kernel Canonical Correlation Analysis that discovers the relationship between how humans tag images (e.g., the order in which words are mentioned) and the relative importance of objects and their layout in the scene. Using this discovered connection, we show how to boost accuracy for novel queries, such that the search results better preserve the aspects a human may find most worth mentioning. We evaluate our approach on three datasets using either keyword tags or natural language descriptions, and quantify results with both ground truth parameters as well as direct tests with human subjects. Our results show clear improvements over approaches that either rely on image features alone, or that use words and image features but ignore the implied importance cues. Overall, our work provides a novel way to incorporate high-level human perception of scenes into visual representations for enhanced image search.

Keywords Image retrieval · Image tags · Multi-modal retrieval · Cross-modal retrieval · Image search · Object recognition · Auto annotation · Kernelized canonical correlation analysis

S.J. Hwang (✉) · K. Grauman
Department of Computer Science, University of Texas at Austin, Austin, TX 78712, USA
e-mail: sjhwang@cs.utexas.edu

K. Grauman
e-mail: grauman@cs.utexas.edu

1 Introduction

Images tagged with human-provided keywords are a valuable source of data, and are increasingly available thanks to community photo sharing sites such as Flickr, collaborative annotation games (von Ahn and Dabbish 2004), online captioned news photo collections, and various labeling projects in the vision community (Russell et al. 2005; Deng et al. 2009). While a user's intentions when tagging may vary, often the keywords reflect the objects and events of significance and can thus be exploited as a loose form of labels and context.

Accordingly, researchers have explored a variety of ways to leverage images with associated text—including learning their correspondence for auto-annotation of regions, objects, and scenes (Duygulu et al. 2002; Barnard et al. 2003; Gupta and Davis 2008; Berg et al. 2004; Li et al. 2009; Hwang and Grauman 2010b), using keyword search for inexpensive dataset creation (Fergus et al. 2005; Li et al. 2007; Schroff et al. 2007; Vijayanarasimhan and Grauman 2008), and building richer image representations based on the two simultaneous "views" for retrieval or clustering (Monay and Gatica-Perez 2003; Hardoon and Shawe-Taylor 2003; Quattoni et al. 2007; Bekkerman and Jeon 2007; Quack et al. 2008; Blaschko and Lampert 2008; Yakhnenko and Honavar 2009; Qi et al. 2009). Such methods have shown that learning with words and images together can yield stronger models.

However, existing approaches largely assume that image tags' value is purely in indicating the presence of certain objects. As such, for retrieval applications, one scores the query results according to the number of shared keywords among the ground truth tags; similarly, for recognition applications, one scores the label predictions according to their per-class accuracy. The problem with this assumption is that

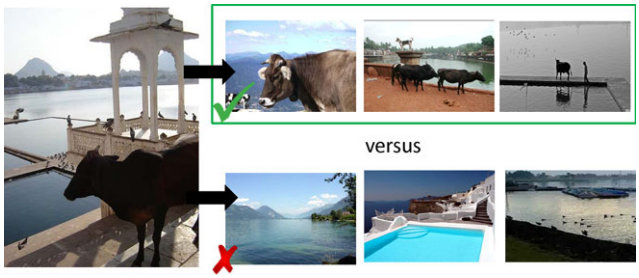


Fig. 1 Illustrative example to convey the role of relative importance in image retrieval. The *left image* is a query, and the *two rows on the right* are two possible sets of retrieval results. Both rows share similar numbers of total objects with the query. However, a human observer may prefer the top row, since those images contain the same most noticeable ("important") objects as the query and roughly agree in terms of those objects' relative prominence. These contrasting retrieval results suggest that a visual index centered solely on the *presence* of high-level tokens (e.g., object categories) may fail to capture their perceived relative *importance.* Our main contribution is to show how to learn representations that respect such cues, and to demonstrate its effectiveness for content-based and cross-modal search

it ignores the relative *importance* of different objects composing a scene, and the impact that this importance can have on a user's perception of relevance. In particular, we consider an object's "importance" in a scene to be directly proportional to *how likely it would be named early on by a human describing the image.*

Figure 1 illustrates the role of perceived importance in image retrieval. Suppose the leftmost image is a query, and the two rows to its right are two sets of retrieved examples. Although both rows of images have roughly similar numbers of shared objects with the query, arguably the top row contains better perceptual matches. Intuitively, we see that the more prominent or noticeable objects of interest from the query are better preserved. This simple example hints that a retrieval system may be lacking if it were to care solely about object presence.

Our goal is to discover the underlying importance cues, and use them to auto-tag or retrieve images with the most prominent objects or those that best define the scene. Interestingly, the abstract nature of the perceived importance in our illustrative example suggests that neither traditional measures of low-level saliency (e.g., Kadir and Brady 2001; Bruce and Tsotsos 2005) nor statistical feature selection techniques (e.g., the well known *tf-idf* weighting typically used in information retrieval Baeza-Yates and Ribeiro-Neto 1999) would be sufficient to capture it.

How can we learn the relative importance of objects and use this knowledge to improve image retrieval? Our approach rests on the assumption that humans name the most prominent or interesting items first when asked to summarize an image. Thus, rather than treat tags as simply a set of names, we consider them as an ordered list conveying useful cues about what is most notable to the human viewer. Specifically, we record a tag-list's nouns, their absolute ordering, and their relative ranking compared to their typical placement across all images. We propose an unsupervised approach based on Kernel Canonical Correlation Analysis (KCCA) (Akaho 2001; Fyfe and Lai 2001; Hardoon et al. 2004) to discover a "semantic space" that captures the relationship between those tag cues and the image content itself, and show how it can be used to more effectively process novel text- or image-based queries or auto-tagging requests.

The semantic space from KCCA is a common representation for both modes of the data constructed such that the correlation between projections from each modality on the same instance (image+tag cues) is maximized. Having learned the desired connections on a batch of human-tagged images, we project all database images into the semantic space. Then, to process a novel keyword-based or image-based query, we project the input onto the semantic space, and find those database items that are nearest. Because we can project data into the learned representation from either visual or text-based features, our method supports three forms of retrieval: content-based image retrieval, keyword-based image search, and automatic tag generation for novel images.

Unlike traditional "appearance only" retrieval systems, we expect our approach to find images that have more semantic (object-level) relevance to the query. Unlike existing approaches that learn from both words and images (including prior uses of KCCA: Hardoon and Shawe-Taylor 2003; Blaschko and Lampert 2008; Yakhnenko and Honavar 2009), we expect to find images that better capture the human's concept of importance and layout. See Fig. 2.

To validate this idea, we perform a series of experiments with three different datasets of images from consumer photo collections and benchmark recognition data. To learn the semantic space representation, we use PASCAL (Everingham et al. 2007) and LabelMe (Russell et al. 2005) images and gather text annotations from workers on Mechanical Turk. Posing tasks to test the different types of retrieval supported by our method, we compare to alternative methods that either use only the visual descriptions, or else use the text during learning but neglect the implicit importance cues. Furthermore, we perform experiments with hundreds of human subjects who are asked to judge the extent to which each method returns images that share the important objects of an image query, thereby directly testing the intended impact our learned representation on perceived importance.

Our main contributions are: (1) an approach to learn the relative importance or configuration of objects that requires no manual supervision beyond gathering tagged images, and (2) experiments demonstrating that the learned semantic space enhances retrieval, as judged by both quantitative measures of object prominence derived from human-provided data as well as experiments with human subjects.

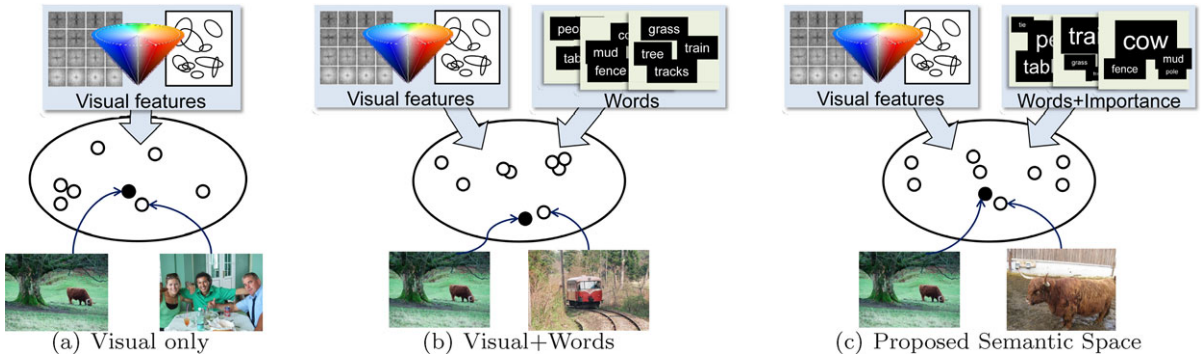


Fig. 2 Main idea and contrasts with alternate representations. (**a**) Most content-based retrieval methods search purely in the visual feature space; this allows one to retrieve similar-looking examples for a query (*bottom left image*/denoted with *black dots*), but can result in semantically irrelevant retrievals (e.g., similar colors and edges, but different objects). (**b**) Some work considers ways to learn a Visual+Words "semantic space" (e.g., with KCCA), which helps narrow retrievals to images with a similar distribution of objects. However, the representation still lacks knowledge as to which objects are more important to a human viewer. (**c**) Our idea is to learn the semantic space using the *order* and *relative ranks* of the human-provided tag-lists. As a result, we retrieve not only images with similar distributions of objects, but more specifically, images where the objects have similar degrees of "importance" as in the query image. For example, the retrieved image on the bottom right, though lacking a 'tree', is a better match due to its focus on the cow

2 Related Work

Our approach most relates to work in image auto-annotation, learning from multiple input channels (text and imagery), and models of human attention and saliency, as we overview in the following. For an overview of the wide area of content-based retrieval, please see reviews by Datta et al. (2008) and Smeulders et al. (2000).

One way to exploit tagged or captioned photos is to recover the correspondence (or "translation") between region descriptors and the keywords that appear alongside them (Duygulu et al. 2002; Berg et al. 2004). The idea is to use the collection of loosely paired data to help resolve ambiguities and discover the mapping that is most consistent. Related approaches develop models for the joint distribution of words and regions (Barnard et al. 2003; Monay and Gatica-Perez 2003; Lavrenko et al. 2003) and scenes (Li et al. 2009). While nearly all such methods focus on the connection between nouns and image regions, some recent work further explores how exploiting predicates (e.g., isAbove(A,B)) can give even stronger cues tying the two data views together (Gupta and Davis 2008). Having learned such a model, one can predict the labels for new image examples. Tackling a similar problem in a different way, Farhadi et al. (2010) show how to generate short descriptive sentences consisting of an object-action-scene triple by evaluating the similarity between a sentence and image.

Such methods are appealing because they capitalize on existing sources of multi-modal data, thereby leveraging a form of "free" (though noisy) labeled instances. Furthermore, their ability to predict text associated with novel examples is valuable for semantically meaningful keyword image search. Our work shares these motivating factors with the above cited techniques. However, beyond learning the association between visual content and the appropriate words, we aim to prioritize their implied importance according to a human viewer. Furthermore, previous methods that first decompose images into regions risk missing the true correspondences due to inevitable flaws in segmentation techniques. In contrast, we consider image-level descriptors together with the complete textual description or list of keywords, which allows us to bypass committing to a bottom-up division into regions.

As such, our approach also relates to methods that learn a new image representation (or similarly, a distance function) that is bolstered by having observed many examples together with relevant text. A number of learning strategies have been explored in the literature along these lines, including variants of metric learning (Qi et al. 2009; Makadia et al. 2008), transfer learning (Quattoni et al. 2007), matrix factorization (Loeff and Farhadi 2008), and random field models (Bekkerman and Jeon 2007). Most relevant to our method, some previous work has specifically considered KCCA for this purpose (Hardoon and Shawe-Taylor 2003; Yakhnenko and Honavar 2009; Blaschko and Lampert 2008). The exact details of the objective functions differ, but in all such cases the underlying idea is to choose parameters in the new visual representation such that labels or similarity constraints derived from the text space are also preserved.

Notably, however, previous methods limit the text description to simple bag-of-words—capturing only "what" was said. In contrast, our key insight is that rich contextual information is available in "how" human-provided tags or descriptions are given, specifically in their order, rank, and proximity. Thus, whereas prior work learning representations with accompanying text associates image content with

an appropriate word or distribution of words (Qi et al. 2009; Makadia et al. 2008; Quattoni et al. 2007; Loeff and Farhadi 2008; Bekkerman and Jeon 2007; Hardoon and Shawe-Taylor 2003; Yakhnenko and Honavar 2009), the semantic space we discover also preserves the relative significance of the objects present. We expect this difference to be most useful in retrieval applications where one wants to access scenes that are perceptually similar though not visually identical, or in auto-tagging applications where very compact focused descriptions are required.

The problem we consider prompts several questions: what objects do people notice most in an image? what do they tag first? and do they generally agree on what is most important? In the image-based ESP "game with a purpose", von Ahn and Dabbish (2004) show that pairs of people who do not know each other can often quickly agree on the right set of words to describe an image. The work gives strong evidence that there is some consistency in how people attempt to most compactly describe an image, which is an underlying assumption of the representation we define.

In other studies, researchers have found that there is a close relationship between the relative semantic importance or saliency of an object, and in what order a user tags the image (Spain and Perona 2008; Einhauser et al. 2008; Elazary and Itti 2008). On the one hand, low-level cues such as those used in interest operators often coincide with objects people find interesting Elazary and Itti (2008). On the other hand, other studies demonstrate that the top-down saliency of recognized objects clearly directs a viewer's attention Einhauser et al. (2008).

Spain and Perona (2008) give a specific definition for *importance*: an object's importance in an image is the probability that it would be named first by a viewer. The authors devise a model for this concept, and demonstrate that one can *predict* the importance of named objects in an image via regression on some intuitive image cues (e.g., scale, saliency, etc.). Their premise of some objects being more important than others motivates our idea, and, like the ESP game results, their empirical findings about agreement across annotators support the feasibility of learning from human-provided tags. However, the system developed in Spain and Perona (2008) assumes that all examples (including the novel test inputs) are already hand-segmented and labeled by object category, which prevents its use for the retrieval applications considered in this work. In contrast, our algorithm uses unsegmented, unlabeled data for both training and testing.

We first began exploring the connection between implicit tag cues and object localization parameters for the sake of object detection (Hwang and Grauman 2010b). In that work we tackle the inverse problem from Spain and Perona: given a novel test image, its tags are used to prime an object detector. There we train with images for which both tags *and* ground truth bounding boxes are available, learning a prior for the scales and positions of the named objects given a set of rank and proximity cues extracted from the tag words. The results indicate that one can exploit the tags as context to know "where to look" for a given object. This in turn yields faster object detection, since we can abandon traditional sliding window search and instead classify windows in the order indicated by the localization prior. Similarly, it allows more accurate object detection, since we can integrate the prior with the detector response to refine the classifier's responses based on image content alone. See Hwang and Grauman (2010b) for details.

In this work we also consider ways to elicit the implied information in textual descriptions of images, and, similar to Spain and Perona (2008) and Hwang and Grauman (2010b), we are interested in the order in which people name objects in a scene. However, our target application is quite distinct. Our goal is to provide more accurate image retrieval by virtue of learning about object importance, not to explicitly detect objects (Hwang and Grauman 2010b) or sort a list of pre-recognized objects (Spain and Perona 2008). Furthermore, whereas we separately learn an object localization priming distribution from tag data in Hwang and Grauman (2010b), here we instead learn a *single shared representation* for the tag and visual data. This is essential to allowing cross-modal retrieval. Finally, our approach requires significantly less supervision than Spain's work or our own work in detection. During training it only needs access to tagged images, and at query time it requires only an image *or* a set of tags (depending on whether one performs image-to-image retrieval or a tag-to-image retrieval).

To our knowledge, no previous work attempts to improve image retrieval based on importance-based semantics automatically gleaned from tagged images. This article expands upon our previous conference publication (Hwang and Grauman 2010a). In this version, we add an entirely new set of experiments on a new dataset consisting of images with natural language captions (Sect. 4 and Figs. 7, 16, and 19), and perform a new human subject experiment to validate that our method can indeed better find the images with the query's important objects as perceived by humans (Sect. 4.5.1, and Figs. 14, 13, and 15). In addition, we add several new figures to illustrate the problem and our algorithm (Figs. 1, 3, 4, and 8), provide additional discussion of the results and related work, and insert new qualitative image retrieval results (Figs. 10 and 11).

3 Approach

Our goal is to provide an image retrieval and auto-tagging system that accounts not only for the objects present in the images, but also their relative significance within the scene.

If it works well, then a user's query should map to images where the most important components of the scene are preserved, and similarly, auto-generated tags should name the most defining objects first. Recall from the introduction that we will consider the more "important" objects to be those that a human would mention earlier when asked to describe the contents of an image.

A naive approach might be to run a bank of object detectors on all database images, and then filter retrieval results according to the predicted objects' sizes and positions. However, such a technique would entail manually specifying global rules about preferred scales and knowing what preset list of detectors is relevant, and would quickly become unaffordable for large databases.

Instead, we propose a lightweight approach that directly learns the implicit cues about importance from human-tagged image data. First, we collect images of the sort an ordinary user might want to search and organize based on content. Then, we obtain tags for these photos via online annotators, whom are simply asked to name the objects in the scene, but with no requirements as to the number or order in which those tags should be provided. We treat the ordered tag words and the image's visual descriptors (color, local features, etc.) as two views stemming from the common semantics of the image content. To learn a representation that exploits both views and allows us to compute similarities *across* the two views (i.e., so that we may support not only image-to-image retrieval, but also tag-to-image retrieval or image-to-tag auto-annotation), we employ Kernel Canonical Correlation Analysis (KCCA). This algorithm essentially learns two sets of basis functions, one per view, such that correlation is maximized for projections of either view of the same instance. Finally, at test time, we project the novel image or tag query onto this learned semantic space, and rank the database images according to their semantic feature similarity.

Our approach makes two key assumptions: (1) people tend to agree about which objects most define a scene, and (2) the significance and prominence of those objects in turn influence the order in which a person provides image tags. Though difficult to state in the absolute, a number of previous studies lend support for both points (von Ahn and Dabbish 2004; Spain and Perona 2008; Einhauser et al. 2008; Elazary and Itti 2008; Tatler et al. 2005; Hwang and Grauman 2010b), as does our own experimental data.

We next define the features and KCCA algorithm (Sects. 3.1 and 3.2), and then describe how we use the combined semantic representation to process three types of queries (Sect. 3.3).

3.1 Tag and Image Features

We examine three types of tag-based features, which together capture the objects present as well as an indirect signal about their inter-relationships in the scene.

3.1.1 Word Frequency

This feature is a traditional bag-of-words. It records which objects are named, and how many times. Supposing V total possible words in the text vocabulary, each tag-list is mapped to an V-dimensional vector

$$W = [w_1, \ldots, w_V], \tag{1}$$

where w_i denotes the number of times the i-th word is mentioned in the list. For tag lists, usually counts are simply 0 or 1. This feature serves to help learn the connection the low-level image features and the objects they refer to, and has been used previously in applications of KCCA (Hardoon and Shawe-Taylor 2003; Yakhnenko and Honavar 2009; Blaschko and Lampert 2008).

The motivation for using word frequency is to learn from both which objects were and were not mentioned. Several objects mentioned together clearly give some scene context, as well as a probable constraint on relative scales. For example, visualize an image that might prompt the description "computer", "mouse", "stapler", versus one that might prompt the description "computer", "chairs", "desk"; we get a cue about the image field of view and the focus point within the scene. In addition, given the tendency to name prominent or large objects in favor of smaller ones Spain and Perona (2008), this feature can help us discover the connection between the "right" visual features when learning the semantic space.

3.1.2 Relative Tag Rank

This feature encodes the relative rank of each word compared to its typical rank:

$$R = [r_1, \ldots, r_V], \tag{2}$$

where r_i denotes the percentile of the i-th word's rank relative to all its previous ranks observed in the training data. Specifically,

$$r_i = 1 - \frac{\sum_{k=1}^{J} w_{ik}}{\sum_{k=1}^{N} w_{ik}}, \tag{3}$$

where w_{ik} is the number of times the i-th word has the k-th absolute rank in the training instances, $J = \min(a_i, N)$, where a_i refers to the average absolute rank of the i-th word in the given image's tag-list.[1] In short, the higher the value of r_i, the more this word tops the list relative to where it

[1] It is the *average* because the same word may appear more than once in some tag-lists. We use $N = 50$ as a cap on the maximum absolute rank, in order to ignore tags that appear late in unusually long lists. On average most have only 5–23 tags; see Sect. 4.2.

typically occurs in any other tag-list; if absent, the percentile is 0.

The motivation for using tag rank as a feature is to capture the correlation between order of mention and objects' prominence within the scene. The sequence of the words given by a human tagging an image is revealing; when forming a compact description of an image we are influenced by objects' scales, centrality within the image, significance, and other attentional cues (Einhauser et al. 2008; Tatler et al. 2005; Elazary and Itti 2008; Wolfe and Horowitz 2004; Spain and Perona 2008). In particular, people exhibit a central fixation bias (Tatler et al. 2005) which means a priori we may tend to mention the central object framed in the shot early in the tag list. The purpose of using the percentile rank rather than the raw rank is to distinguish those words (objects) that may be atypically prominent.

3.1.3 Absolute Tag Rank

This feature encodes the absolute rank of each word:

$$A = \left[\frac{1}{\log_2(1+a_1)}, \ldots, \frac{1}{\log_2(1+a_V)}\right], \tag{4}$$

where a_i denotes the average absolute rank of the i-th word in the tag-list. Note that A will be 1 if the object is mentioned first, and will drop exponentially towards 0 for lower ranked or absent words.

This feature has a similar motivation to relative rank defined above. However, in contrast to the relative rank, it more directly captures the importance of each object compared to the others in the *same* scene. We expect the relative rank to be more indicative of the scale and position of each object, while we expect the absolute rank to be more indicative of its semantic importance for that image instance.

For example, suppose an image has the following tag list: {car, stop sign, person, tree}. Here, 'stop sign' and 'person' occupy the second and third place in the list, respectively, and thus will have higher absolute rank than 'tree', which appears fourth. However, if the person and stop sign tags typically occur first in the training data, and tree typically occurs seventh, then the relative rank of tree would be higher than the relative rank for the other two. The relative rank essentially normalizes for the inherent semantic importance of each object (i.e., 'person' is often tagged early no matter what), and so our two variants of rank give slightly different information.

3.1.4 Visual Descriptors

We extract three visual features for every image: a Gist descriptor, a color histogram, and a bag-of-visual-words histogram.

The Gist is a 512-dimensional vector recording the pooled steerable filter responses within a grid of spatial cells across the image (Torralba 2003). It captures the total scene structure.

The color histograms capture the global distribution of colors, and can also be useful to describe overall scene type (e.g., green vegetation, gray urban areas, multi-colored indoor scenes, etc.). We use 64-dimensional HSV color histograms, with 8, 4, and 2 bins for hue, saturation, and value, following Blaschko and Lampert (2008).

The bag-of-words (BOW) summarizes the frequency with which each of a set of prototypical local appearance patches occurs; we use DoG interest point selection and SIFT descriptors (Lowe 2004), and form 200 words with k-means. These local features are useful to capture the appearance of component objects, without the spatial rigidity of Gist.

3.2 Kernel Canonical Correlation Analysis

Given the two views of the data, we are ready to construct their common representation. Canonical Correlation Analysis (CCA) uses data consisting of paired views to simultaneously find projections from each feature space such that correlation between projected features originating from the same instance is maximized (Hotelling 1936). The algorithm learns two semantic projection bases, one per descriptor type.

Formally, given samples of paired data $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, where each $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ denote the two views, the goal is to select directions $w_x \in \mathbb{R}^m$ and $w_y \in \mathbb{R}^n$ so as to maximize the canonical correlation:

$$\begin{aligned} w_x^*, w_y^* &= \arg\max_{w_x, w_y} \frac{\hat{E}[\langle x, w_x\rangle\langle y, w_y\rangle]}{\sqrt{\hat{E}[\langle x, w_x\rangle^2]\hat{E}[\langle y, w_y\rangle^2]}} \\ &= \arg\max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x w_y^T C_{yy} w_y}}, \end{aligned} \tag{5}$$

where $\hat{E}$ denotes the empirical expectation, C_{xy} denotes the between-sets covariance matrix, and C_{xx} and C_{yy} denote the auto-covariance matrices for x and y data, respectively. Note that in our case, the x view is the visual cue, and the y view is the tag-list cue. The solution can be found via a generalized eigenvalue problem. The CCA method has often been used in cross-language information retrieval, where one queries a document in one language to retrieve relevant documents in another language (Li and Shawe-Taylor 2006).

Kernel CCA is a kernelized version of CCA. Given kernel functions for either feature space:

$$\begin{aligned} k_x(x_i, x_j) &= \phi_x(x_i)^T \phi_x(x_j), \\ k_y(y_i, y_j) &= \phi_y(y_i)^T \phi_y(y_j), \end{aligned} \tag{6}$$

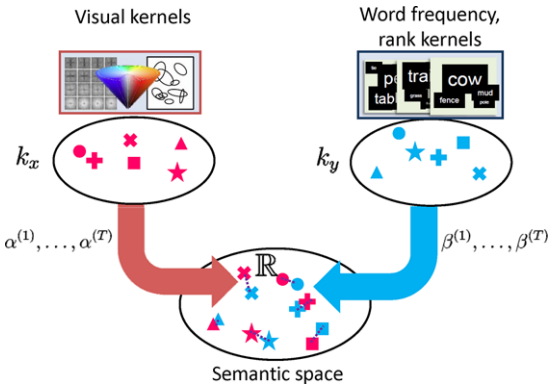


Fig. 3 Schematic illustrating Kernel Canonical Correlation Analysis (KCCA) for our two-view data consisting of visual and tag-based descriptors. The projections learned from the paired training instances map data from either space into a common "semantic space", such that the correlation between the two views is maximized. Here the symbol shapes (*star*, *square*, etc.) denote a single instance

one seeks projection vectors in the kernels' implicit feature spaces, which may only be accessed through kernel function evaluations. The solution for w_x and w_y must lie in the span of the N training instances $\phi_x(x_i)$ and $\phi_y(y_i)$:

$$w_x = \sum_i \alpha_i \phi_x(x_i),$$
$$w_y = \sum_i \beta_i \phi_y(y_i), \tag{7}$$

where $1 \leq i \leq N$.

The objective in the kernelized form is thus to identify the weights $\alpha, \beta \in \mathbb{R}^N$ that maximize

$$\alpha^*, \beta^* = \arg\max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \beta^T K_y^2 \beta}}, \tag{8}$$

where K_x and K_y denote the $N \times N$ kernel matrices over a sample of N pairs. This can also be formulated as an eigenvalue problem (with additional regularization modifying the above to avoid degenerate solutions), and the top T eigenvectors yield a series of bases $(\alpha^{(1)}, \beta^{(1)}), \ldots, (\alpha^{(T)}, \beta^{(T)})$ with which to compute the T-dimensional projections for an input x or y. See Hardoon et al. (2004) for more details.

We use χ^2 kernels for all component visual and tag-based features:

$$K_{\chi^2}(h_i, h_j) = \exp\left(-\frac{1}{2\Omega}\sum_{k=1}^{d}\frac{(h_i(k) - h_j(k))^2}{h_i(k) + h_j(k)}\right), \tag{9}$$

where Ω denotes the mean of the χ^2 distances among the training examples, and d denotes the dimensionality of the descriptor. We average the K_{χ^2} kernels for the image features, i.e., $k_x(x_i, x_j)$ is an average of the respective K_{χ^2} kernels for Gist, color, and BOW. For the tag-based kernel $k_y(y_i, y_j)$, we average the K_{χ^2} kernels for our tag features R and A. We use only W for the Visual+Word baseline (see Sect. 4).

Figure 3 illustrates how the learned KCCA kernel-space projections map data from either the visual descriptor space or the tag-based feature space into a common semantic space. Note that once this space has been learned, we can project novel data from *either* view independently into the shared space.

3.3 Processing Novel Queries

KCCA provides a common representation for the visual and tag features. Given a novel visual input q_x, we project onto a single basis specified by α as:

$$w_x^T \phi_x(q_x) = \sum_{i=1}^{N} \alpha_i \phi_x(x_i)^T \phi_x(q_x)$$
$$= \sum_{i=1}^{N} \alpha_i k_x(x_i, q_x). \tag{10}$$

Thus projecting a novel image descriptor requires evaluating the kernel function on it and those of the N training points which have nonzero weights. Similarly, the projection of a novel tag-list input q_y is given by:

$$w_y^T \phi_y(q_y) = \sum_{i=1}^{N} \beta_i k_y(y_i, q_y). \tag{11}$$

The final projection of q_x or q_y onto the T-dimensional semantic space is formed by the vector of these values for $\alpha^{(1)}, \ldots, \alpha^{(T)}$ or $\beta^{(1)}, \ldots, \beta^{(T)}$, respectively.

After projecting all database images onto this learned semantic space, there are three tasks we can perform, as we outline in the following and depict in Fig. 4.

3.3.1 Image-to-Image Retrieval

The first task is content-based image retrieval. Given a novel query image, we use its image features only $(\phi_x(x))$ to project onto the semantic space, and then sort all database images relative to it based on their normalized correlation in that space. Figure 4(a) illustrates the retrieval process.

Compared to the results we would expect using the database's image features *alone* to rank the data, the results should be favorably skewed towards showing scenes with similarly relevant objects when using our approach.

3.3.2 Tag-to-Image Retrieval

The second task is keyword-based image retrieval. Given a novel tag-list query $(\phi_y(y))$, we project onto the semantic

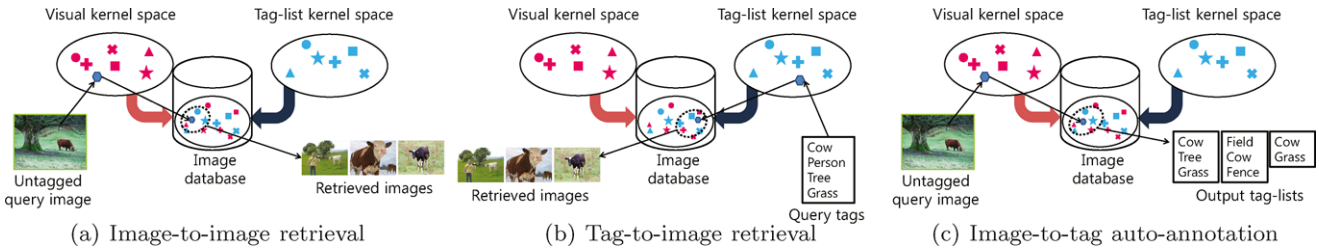


Fig. 4 Illustration of the three types of query tasks performed by our approach. See text for details

space, and again sort the database images by correlation. Figure 4(b) illustrates the retrieval process.

Compared to traditional keyword-based image search, we can now expect to retrieve images where not only are objects shared with the query, but they are also of similar relative importance.

3.3.3 Image-to-Tag Auto-Annotation

The third task is auto-annotation. Given a novel query image, we project onto the semantic space, identify its K nearest examples among those that are tagged, and merge their keywords to create an output tag-list. Specifically, we merge all tag occurrences on the K retrieved tag-lists, and then reorder the tags by their frequency to produce the output. Figure 4(c) illustrates the auto-annotation process.

Compared to existing approaches that auto-annotate by predicting labels for each blob in the image, this strategy attempts to provide the most *important* tags based on all available image features.

3.3.4 Computational Cost

The primary offline training cost is solving the eigenvalue problem for KCCA. Note that the image database may be a superset of the N images used to train KCCA, and need not be fully tagged, since any novel untagged image can be projected onto the learned semantic space.

To retrieve the nearest neighbors given a query for any of the three query types outlined above, we simply compute a linear scan of the database. A faster sub-linear time implementation could also easily be incorporated, for example, with Kernelized Locality Sensitive Hashing (Kulis and Grauman 2009).

4 Experimental Results

In this section, we apply our method for each of the three scenarios outlined above. The primary goal of our experiments is to demonstrate that retrieval quality and auto-annotation accuracy can be enhanced by accounting for the relative importance of objects. We compare our algorithm to the most relevant baseline approaches—including an alternative KCCA baseline that uses unordered keywords—and show results on three different challenging datasets.

4.1 Baselines

We compare our approach to three baselines:

- *Visual-only*, which ranks images relative to an image query according to their visual descriptors only,
- *Word-only*, which ranks tagged images relative to a keyword query according to tag similarity only,
- *Word+Visual*, a strong baseline that builds a KCCA semantic space using the image cues plus a bag-of-keywords W. This approach is very similar to multi-view retrieval frameworks developed in previous work (Yakhnenko and Honavar 2009; Hardoon and Shawe-Taylor 2003), and therefore is a good representative of how existing techniques would integrate words when learning a visual representation.

While a wealth of image retrieval techniques have been explored in the literature (Datta et al. 2008; Smeulders et al. 2000), to our knowledge none is concerned with retrieving images that share the most important objects with a query, nor do any previous methods aim to auto-tag in the appropriate order so as to mimic human-provided tag-lists. This is the key novelty of our problem definition and approach. Thus, comparisons to numbers reported in the literature for generic CBIR tasks are not suited to support our claims, and the baselines listed above are the most important natural alternatives to test in order to validate our idea.

To allow the most informative comparisons, for all methods we use the exact same visual features and kernels. Our method uses R and A together with the visual cues.

4.2 Datasets

We consider three datasets: (1) the PASCAL VOC 2007 images (Everingham et al. 2007) with keyword tags collected on Amazon's Mechanical Turk, (2) LabelMe images (Russell et al. 2005) with keyword tags collected on Mechanical

Fig. 5 Example PASCAL images with keyword tags collected on Mechanical Turk. Note how the relative importance of the objects affects the order of the tags, as in the fourth example, where the image depicts an atypical scene

Fig. 6 Example LabelMe images with keyword tags collected on Mechanical Turk. Unlike the PASCAL images, the images in this dataset are mostly scenes with no clear central object; here the lower level saliency often plays a more crucial role, such as in the third example above

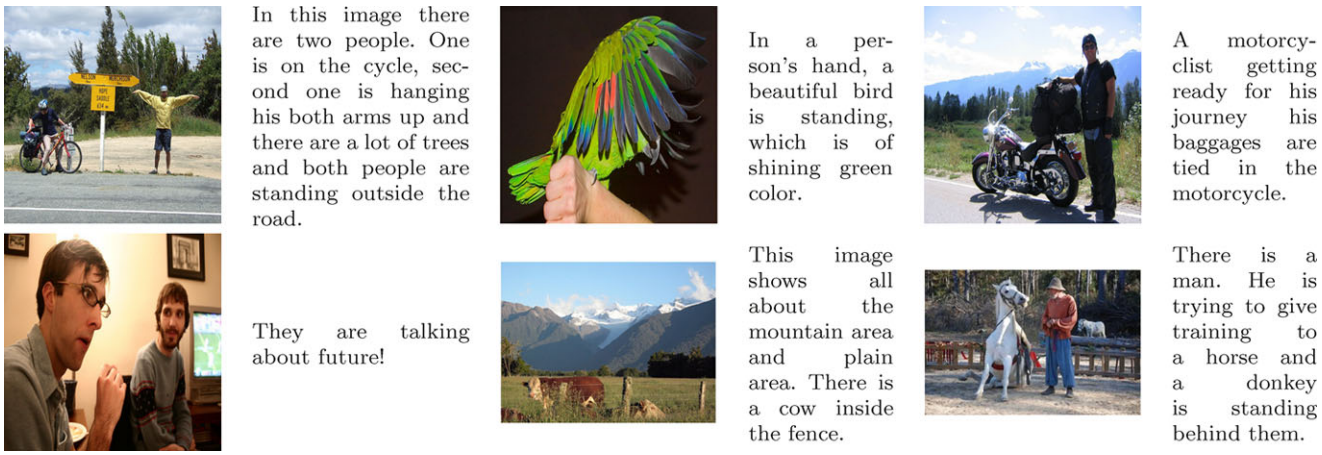


Fig. 7 Example PASCAL images with natural language descriptions collected on Mechanical Turk

Turk, and (3) the PASCAL VOC 2007 images with natural language sentences collected on Mechanical Turk. See Figs. 5 through 7 for examples from each dataset. The PASCAL images in particular are a realistic representation of the sort of images an ordinary user might want to search and organize based on content, since they originate from user-uploaded content on Flickr.

For the first PASCAL dataset, we use the VOC train and test splits as our database (5011 images) and query examples (4952 images). We use the tags we collected in Hwang

and Grauman (2010b) from 758 workers on Mechanical Turk. To collect those tags, we posted each image online with a nearby textbox, and the anonymous workers were instructed to name the objects or items in the image. For quality control, we disabled the textbox until the image had been viewed by the tagger for 7 seconds, and required him/her to submit after 30 seconds had passed. We refine the resulting tags using the LabelMe toolbox, correcting spelling errors and resolving synonyms. PASCAL contains fairly object-centric instances, with relatively few tagged objects per image—5.5 tags on average among $V = 399$ total words. See Fig. 5 for example images.

The LabelMe images contain broader scenes (often offices and streets), with many more objects per image. We use the 3825 images compiled in Hwang and Grauman (2010b) for which there are at least 10 tags (on average each has 23). We report results for five 50–50 random database-query splits. See Fig. 6 for example images.

To test the PASCAL images with natural language descriptions, we use Mechanical Turk to gather sentences from workers on a randomly selected subset of 500 images. In contrast to the above tag collection, in this case we instruct the workers to give a detailed explanation of the scene without any other restrictions. We manually correct misspellings and resolve synonyms using the same tools as above; further, we remove all articles, pronouns, and prepositions from the vocabulary in order to focus the representation on key nouns, verbs, and adjectives. We report results for ten 60–40 random database-query splits. See Fig. 7 for example images. The tag-list data and natural language caption data we collected are publicly available.[2]

4.3 Implementation Details

We use the KCCA code provided by Hardoon et al. (2004).[3] We fix the learning rate $\mu = 0.5$ and set the regularization parameter by maximizing the difference in spectrums of the kernels for each view, between samples of actual and random image-tag pairs. Specifically, we set the regularization parameters κ_x and κ_y for the image and tag views, respectively, by maximizing the following:

$$\kappa_i^* = \arg\max_{\kappa_i} \|\lambda_{R_i}(\kappa_i) - \lambda_i(\kappa_i)\|_2, \tag{12}$$

where λ_i is the spectrum of the kernel associated for each view, and λ_{R_i} is the spectrum of the same kernel matrix with random permutation. See (Hardoon et al. 2004) for details.

We fix $T = 20$ for all KCCA projections. In initial informal tests, we did not find the overall results to be very sensitive to the semantic space dimensionality.

[2] http://vision.cs.utexas.edu/projects/importance.

[3] http://www.davidroihardoon.com/Research/Code.html.

4.4 Evaluation Metrics

We score the methods using the Normalized Discounted Cumulative Gain at top k (NDCG@k), a measure commonly used in information retrieval (Jarvelin and Kekalainen 2002). It reflects how well a computed ranking agrees with the ideal (ground truth) ranking, and more strongly emphasizes the accuracy of the higher ranked items. It is defined as:

$$\text{NDCG@}k = \frac{1}{Z}\sum_{p=1}^{k}\frac{2^{s(p)}-1}{\log(1+p)}, \tag{13}$$

where Z is a query-specific normalization term, p cycles over the top k ranks, and $s(p)$ denotes the *reward* at rank position p. The score ranges from 0 to 1, where 1 indicates perfect agreement.

We define reward functions for two variants of "ideal" rankings, both intended to reveal how well the retrievals exhibit the query's important objects. The first is *object counts and scales*, where the ideal ranking would sort the images based on the correlation of the presence and relative scales of all objects named in the query's ground truth tag-list. This reward function is defined as:

$$s(p) = \frac{1}{2}\left(\frac{\langle S_p, S_q\rangle}{\|S_p\|\,\|S_q\|} + \frac{\langle W_p, W_q\rangle}{\|W_p\|\,\|W_q\|}\right), \tag{14}$$

where S_p and W_p are V-dimensional vectors recording the scale (normalized by image size) and count of each object within the p-th retrieved image, respectively, and S_q and W_q are the same for the query.

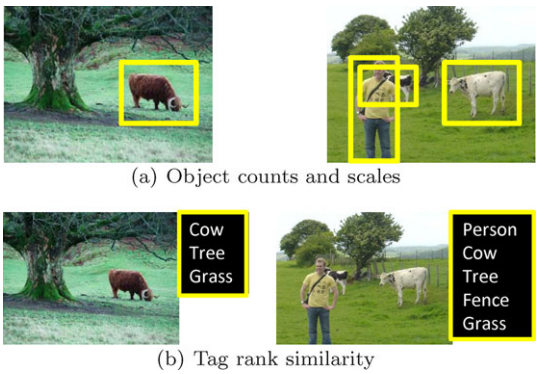
The second reward function conveys *tag rank similarity*. For this scoring, the ideal ranking would sort the images based on the agreement between the (withheld) ground truth ordered tag-list features:

$$s(p) = \frac{1}{2}\left(\frac{\langle R_p, R_q\rangle}{\|R_p\|\,\|R_q\|} + \frac{\langle A_p, A_q\rangle}{\|A_p\|\,\|A_q\|}\right), \tag{15}$$

where the subscripts again refer to the p-th ranked examples and the query image. Note that this metric will reveal to what extent the retrieval results are ranked in a way human taggers would similarly tag them with ordered importance.

We stress that in either reward function, the evaluation relies on ground truth information that remains hidden to our algorithm when processing the novel queries.

The two metrics offer complementary insight into the retrieval results. While the object counts and scales directly uses the objective presence and prominence of each object in a pair of images to gauge their desired similarity, the tag similarity scoring directly uses the more subjective human-provided ordered lists to gauge the desired similarity. For example, consider the images in Fig. 8. The object counts and



(a) Object counts and scales

(b) Tag rank similarity

Fig. 8 The two reward functions used to score NDCG@k account for instances' agreement according to either the ground truth object presence and relative scales (*top*), or the ground truth tag ordering (*bottom*)

scales scoring (a) would consider the cow instances to be very similar, but the lack of a person in the first view would be penalized. The tag rank similarity scoring (b) would penalize the differences more because of the differing 'cow' ranks assigned in either instance, as well as the additional objects mentioned in the second instance.

Finally, as a third measure of accuracy, we ask human subjects to subjectively evaluate results for the image-to-image retrieval task. This allows us to directly study to what extent our method and the baselines retrieve images that a human viewer perceives as capturing the most important elements. We discuss this metric in more detail in the latter part of Sect. 4.5.1.

4.5 Image-to-Image Retrieval Results

First we show that our approach better retrieves images matching the important objects in a query. We first analyze performance when learning the semantic space with tagged images (Sect. 4.5.1), and then we show results when learning with images and full sentence captions (Sect. 4.5.2).

4.5.1 Learning with Tagged Images

For results learning from tagged data, we first quantify accuracy in terms of the information retrieval metric NDCG, and then consider a subjective human evaluation.

Performance evaluated with NDCG@k metrics Figure 9 shows the results for our method and the two baselines on the tagged PASCAL and LabelMe datasets. While the Word+Visual semantic space improves over the Visual-only retrievals, our method outperforms both methods, in terms of the object presence and scales (left plots), and the tag-list agreement (right plots).

Our method's gains are quite significant on the challenging PASCAL data, which makes sense given that the images contain a clearer mix of more and less prominent objects, and there is wider variance in a given category's importance across images. Looking at the $k = 30$ top-ranked images (the number of images one might fit in a Web page search return), we see our method yields a 39% improvement in NDCG over the Visual-only result, and about a 17% gain over the Word+Visual semantic space.

In comparison, LabelMe's broad scenes make it less apparent which objects are most important, and instances of the same scene type contain less variance in composition (e.g., office images tend to have the computer in a similarly prominent role). This allows the traditional semantic space using unordered words to be almost as effective—especially under the tag rank similarity scoring (see bottom two plots in Fig. 9). In other words, we do not expect our approach to offer an advantage when the objects play a fixed role in the scene, not varying in their importance. The example retrievals in Fig. 11 also illustrate this point.

Looking more closely at the plots, we see that the NDCG@k values for the object counts and scales reward function (left two plots) show a decreasing trend. What happens is that many of the lower ranked examples (higher values of k) will have a score of zero for each of the test examples, whereas the ideal ranking method does not. This is because all three methods use background visual information, yet the PASCAL dataset is only labeled for a subset of the foreground objects. On the other hand, for the tag rank similarity reward function (right two plots) there is a gradual increase in NDCG. This is because the tags provided by MTurk workers on the PASCAL images are generally more complete than the PASCAL bounding boxes themselves, and so all methods accumulate more of the relevant database instances for increasing k.

In all plots, we notice that for very small values of k the three methods give similar results, since the first few matches tend to be very close visually *and* semantically. We illustrate this effect further in the qualitative results below.

Figures 10 and 11 show example image retrieval results that illustrate our method's advantages on either dataset. Overall, we see in these examples how the Visual-only baseline can retrieve image instances that closely match the holistic appearance of the query, whereas the Words+Visual baseline often enhances those results to focus on a similar distribution of objects as present in the query. In contrast, our technique tends to retrieve images that further prioritize the most prominent objects from the query.

For example, in the PASCAL results in Fig. 10, the third row shows a query that is a black and white image of a bus on the street. While other objects also appear in the query, the bus is large and dominates the semantics of the image. The Visual-only baseline retrieves images with similar black and white texture patterns, but are otherwise lacking in semantic agreement.[4] The Word+Visual baseline fo-

[4]Recall that color is one of the visual features.

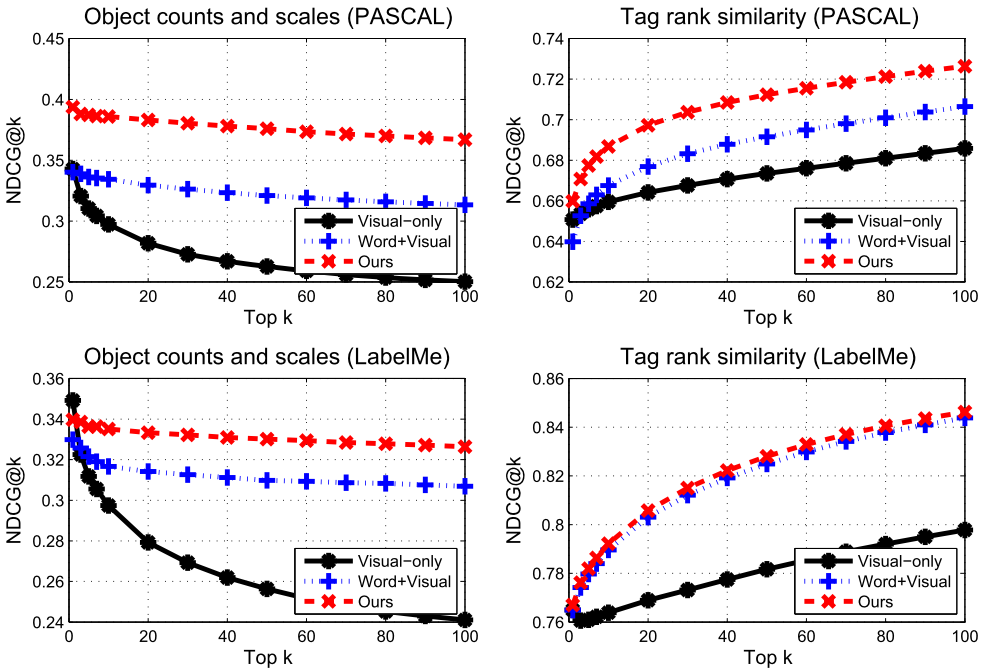


Fig. 9 Image-to-image retrieval results for the tagged PASCAL (*top row*) and LabelMe (*bottom row*) datasets. *Left two plots* use the object counts and scales reward function, while the *right two plots* use the tag rank similarity reward function. Higher curves are better. By modeling importance cues from the tag-lists when building the semantic space, our method outperforms both a method using image features alone, as well as a semantic space that looks only at unordered keywords

cuses more on the contents of the image; however, since it has no cues about the relative importance of the objects, it retrieves images of other street scenes that have buildings or cars as main objects. In contrast, with our method, the mapping to the semantic space has been learned so as to favor the concept 'bus' over other concepts relevant to this scene type, and thus we retrieve other images with a prominent bus. Note also the variety of viewpoints in those examples.

Another useful aspect of our method visible in the image-to-image retrieval results is that it reduces the confusion between classes. For example, in the fifth row of Fig. 10, the Word+Visual baseline retrieves images containing bicycles, instead of birds as shown in the query. This happens because all objects effectively get equal "weight" when the baseline learns the mapping to the semantic space. 'Birds+forest' and 'bicycle+forest' are therefore mapped to a point in the semantic space near 'forest' using this equal weight scheme, introducing confusion between the features that describe birds and bicycle. This problem is accentuated by the fact that in many cases the background features (i.e., for 'forest') dominate the visual representation and have less intra-class variance, while the foreground objects can be more variable. In contrast, during training our method learns to map 'birds+forest' to a point in the semantic space where the concept bicycle does not dominate, thus reducing confusion at test time. Thus, in some sense, our method behaves like distance metric learning: we learn which semantic dimensions are more important than others for the tagged images.

Performance evaluated with human opinions The quantitative analysis so far shows a clear benefit from using our approach, and in particular the tag agreement reward function directly validates that its retrieved images better reflect the things a human observer would mention first when looking at the image. Next, to further assess our accuracy, we perform an experiment with human subjects. The goal is to let the human subjects judge the different retrieval results based solely on their opinion of which contain the same important objects as a query image.

To this end, we performed a leave-one-out test with 5000 total PASCAL image queries, in which we present the human judges with the query image plus the top seven images retrieved by each method. Figure 12 shows the interface for the task. The human subject is asked to select those images that contain the same important objects as the query image. Specifically, we give the following simple instructions: "First look at the topmost query image. Then select those images that contain the most important objects seen in the query." We set up this task such that the same human judge considers results from all methods at once for a given query, yet is given no indication that the results stem from different methods. This way we know that the results are judged side-by-side; if one image in the set is selected but another is not, we know the first was found to be more importance-preserving than the second. We stress that the MTurk workers have no information about which image comes from which method, nor any background on what we are testing beyond the simple instructions shown in Fig. 12.

In order to access a large number of unbiased human subjects with no knowledge of our method or purpose, we posted the tasks on Mechanical Turk. We took two steps to ensure quality control. First, we disabled the "submit" button on the interface until at least 15 seconds had passed,

Fig. 10 Example image-to-image retrievals for our method and the baselines on the tagged PASCAL dataset; *leftmost image* is query, *three rows* show top ranked results per method. While a baseline that builds the semantic space from Words+Visual features can often retrieve images with an object set overlapping the query's, ours often better captures the important objects that perceptually define the scene

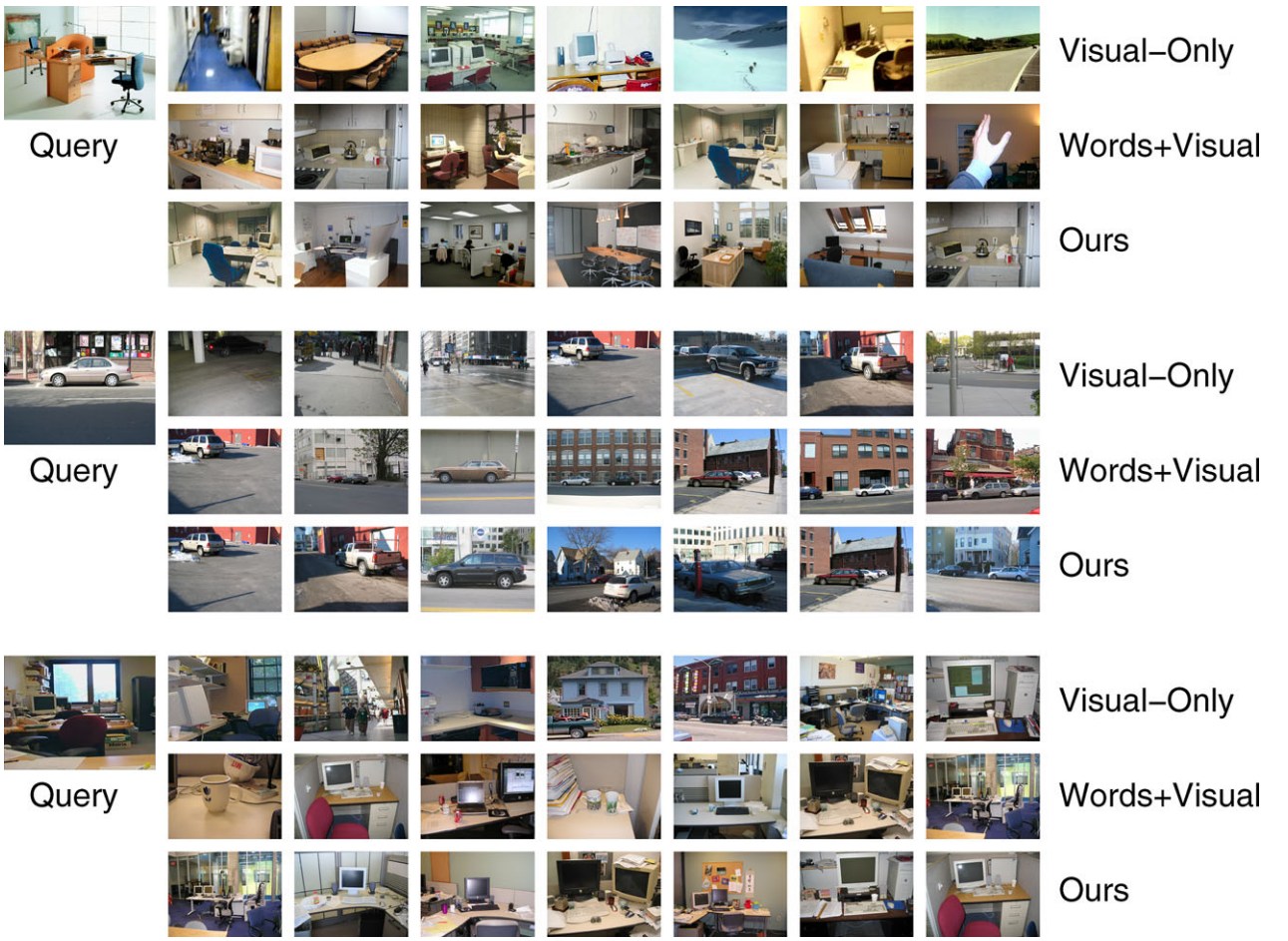


Fig. 11 Example image-to-image retrievals for our method and the baselines on the LabelMe dataset; *leftmost image* is query, *three rows show* top ranked results per method. Our method often better captures the important objects that perceptually define the scene; since most LabelMe images are less object-centric (compared to PASCAL), our method tends to retrieve images with similar scene layouts

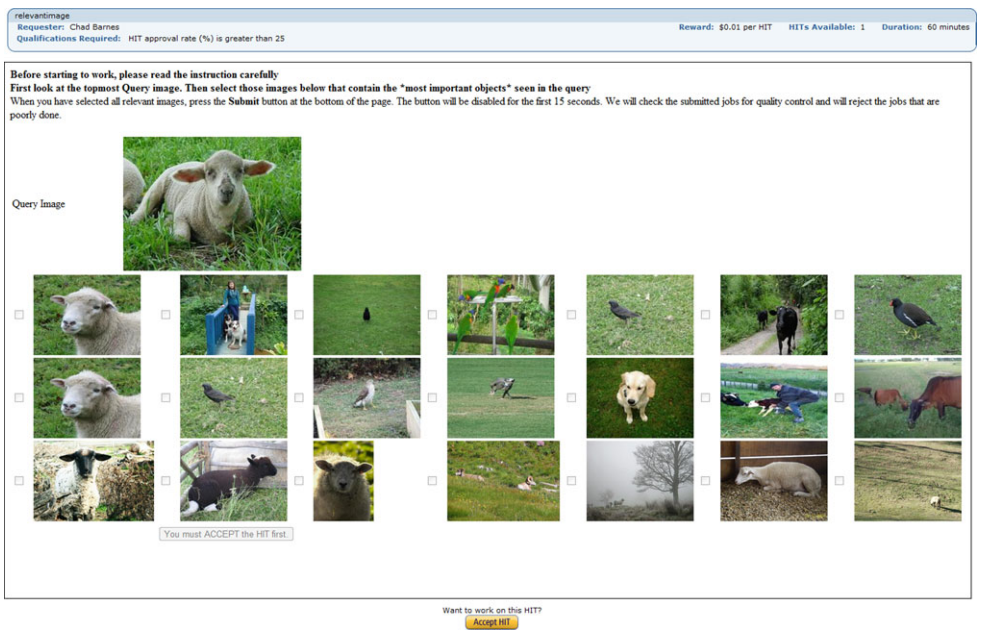


Fig. 12 Mechanical Turk interface for human evaluation of retrieval relevance. The top seven retrieval results are shown for each test query, and the workers are asked to select any image(s) they find to share the "most important objects" with the query by clicking on the checkboxes

in an effort to stop workers from speeding through the task. Notice that workers could select as many or as few images as they liked in each query case. Second, we collected five redundant responses for each query. After running the collection, we obtained results from 323 unique Mechanical Turk workers.

Despite the subjective nature of the task, we find the agreement between the five workers working on the same example is moderately high. Specifically, using Fleiss' kappa statistical test for agreement between multiple raters (Fleiss 1971), we find the agreement score to be $\kappa = 0.4402$. To give some context for this score, complete agreement yields $\kappa = 1$, whereas no agreement other than what is expected by chance yields $\kappa \leq 0$. To refine the human responses to form a ground truth rating, we treat only those images unanimously selected by all five annotators as being relevant.

Analyzing the resulting human subject data, we find that for a large portion of the test examples, no images are selected as relevant (from any method). We suspect this is because the human subjects often found retrieved images sharing similar objects irrelevant when the finer-grained "labels" of the objects differed, as they often do in PASCAL images. For example, while in PASCAL terms any instance of "dog" is related, to the human viewers, a chihuahua or German Shepherd may seem different enough to not share importance/relevance. Thus, this data from humans serves as a useful complementary study to our evaluation with NDCG@k presented above.

Figure 13 shows the results, for all three methods and all 5000 queries. The curves record the average cumulative relevance score per query. For clarity, we sort the queries in the plot by the *total* relevance score obtained over all three methods—that is, the total number of images among all 21 candidates that were unanimously selected by the human subjects as relevant. We see that overall human observers report our results to more often depict the important objects in the query image. This is an important result. It shows that our approach does indeed better capture perceived importance, since that was exactly what the human subjects were instructed to judge. Figure 15(a) shows representative examples that illustrate where our method has an advantage.

At the same time, however, for those queries that yielded the highest number of relevant images across all methods (i.e., the leftmost part of the plot in Fig. 13), we see that the baselines are as good or slightly better than our method. Upon examination, we find that this is likely due to a similar effect we observed in the NDCG results on LabelMe: when there are near-exact visual matches, the retrieval is effectively handled by visual matching alone. Figure 15(b) illustrates such a case, where all methods can obtain very close visual matches for the equestrian query, and our method does not have an advantage by representing the objects' relative importance. Furthermore, we observe that when the 21

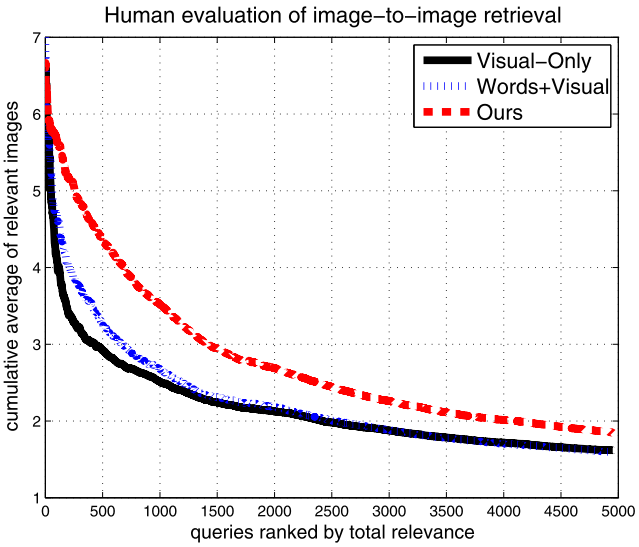


Fig. 13 Average relevance scores across all queries, as judged by human raters on image-to-image retrieval results. The x-axis indexes the 5000 total query examples, where queries are sorted by the total number of relevant images retrieved among the union of the three methods' results. The y-axis measures the cumulative average of the number of relevant images returned in the top 7 results. (Max score on any query is 7, and higher curves are better.) Human observers find that our method's results more often depict the important objects in the query image. However, for query images that have few or no relevant matches in the database, all three method do similarly (right end of the plot)

candidates contain a mix of very close visual matches together with more distant but still semantically relevant images (e.g., the same main object but from a different viewpoint), the human judges tend to favor marking the near-exact matches as relevant. This also reduces our method's impact for the top few "easiest" queries. Finally, there are some difficult queries for which no method retrieves very good results. In such cases we observe the limitations of the low-level visual descriptors, which may be insufficient for some unusual views of objects (see Fig. 15(c) for an example).

Whereas Fig. 13 summarizes the results for all queries, Fig. 14 breaks out the relevance score per method individually for the top 120 queries sorted by total relevance. To help interpret this data, we note that the average number of relevant images obtained by our method is 5.64, which corresponds to 81% precision among the top 7 retrievals. In comparison, the Visual-only baseline averages only 4.12 relevant examples, or 59% precision, while Word+Visual yields a slightly better 4.70, or 67% precision. This plot also reveals that for those queries where Visual-only does better, both the Word+Visual baseline and our method obtain a fairly low score (look at cases where 'x' is high). We again attribute this to cases where there are near-exact matches available for the query. The human evaluators are affected by this, and become more prone to discard semantically similar but visually distinct results. Thus, while overall our method

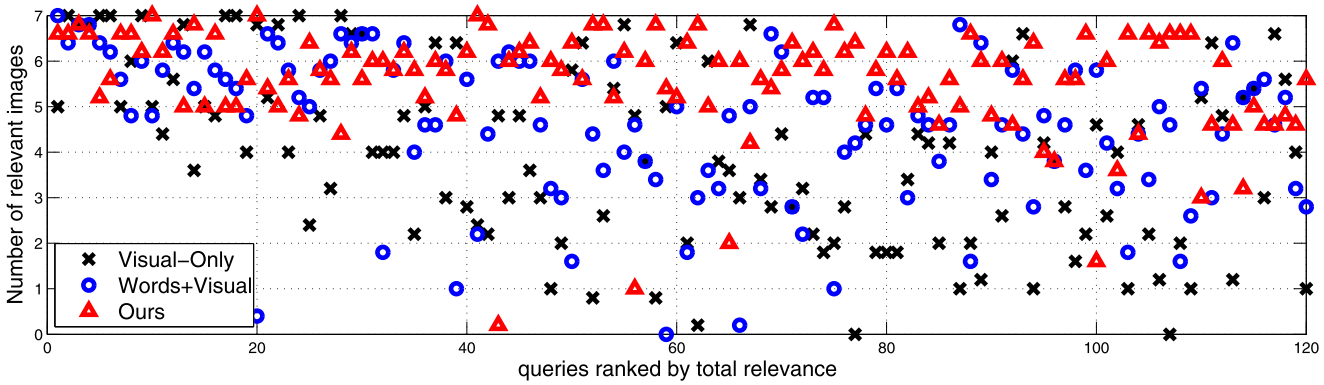


Fig. 14 Relevance scores per query according to human judges, for image-to-image retrieval on PASCAL dataset. Our method shows a clear advantage over the other methods. See text for details

is most consistent in quality, this phenomenon naturally suggests learning when to switch between the learned semantic space and raw visual features; we leave this idea for future work.

4.5.2 Learning with Captioned Images

Next we consider the same experimental image-to-image retrieval setting, but where the semantic space is learned from images that have natural language captions. As depicted in the data-set examples shown in Fig. 7, the semantics in the accompanying text will be much more diverse than the simple tags in this setting. Notably, the sentences contain adjectives that directly describe attributes of objects in the scenes, and verbs noting the ongoing action. This suggests our representation could benefit from the more thorough descriptions. On the other hand, the descriptions are more varied in length compared to the typical tag lists. Some people are quite succinct, others are verbose. Some veer away from pure descriptiveness, and comment on what is imagined about the image, not merely what is visible. Such factors may make it more difficult to learn a clear association with importance. At the same time, we expect that the positions of words in the sentence are less immediately indicative of object importance, since the word order is determined not only by importance but also by grammatical structure in the sentence.

However, in examining the MTurk sentences we collected, we do observe some correlation. First, words that appear earlier in the sentence are likely the subject and thus define the scene to the viewer. Second, the earlier sentences provided for an image tend to comment on the more noticeable objects (subject or otherwise). We also observe that people will invert the subject/object at times in order to emphasize some object in the scene.

Figure 16 shows the results for our method and the two baselines on the PASCAL images and sentence descriptions. As with the tagged dataset, our method again outperforms both the Word+Visual and Visual-only methods. Looking at the $k = 30$ top-ranked images, we see our method yields a 13% improvement in NDCG over the Visual-only result, and about a 5% gain over the Word+Visual semantic space. This shows that our method can generalize to free text sentences.

However, we do find that the absolute improvement in retrieval accuracy is smaller than it is in the above tagged-image experiments. This is likely because the complex structure of natural language sentences reduces the correlation between the objects' importance and their order of mention. In addition, we see narrower gains under the tag rank similarity metric compared to the object scales metric (see right plot in Fig. 16); this is likely because larger differences in the retrieval evaluation using the ground truth bounding boxes can correspond to smaller changes in the sentence configurations. Overall, considering the small size of the training set, and the fact that no further postprocessing is done on the syntax, this is an encouraging result. We leave as future work to explore more elaborate text-based representations that exploit knowledge about grammatical structures.

4.6 Tag-to-Image Retrieval Results

All of the results presented thus far handle the content-based search problem. Next we consider the tag-to-image cross-modal retrieval setting, in which a person queries for images with keywords.

Figure 17 shows the results when we query with a human-provided ordered list of keywords, and return relevant images from the database. Again, the learned semantic space allows us to find the relevant content, this time where the objects emphasized by the human are more likely to be prominent in the results. Our approach makes dramatic improvements over the baselines—31% better in NDCG than the Words+Visual baseline for $k = 30$ on the PASCAL. (Note that we omit scoring with tag rank similarity reward function for this case, since it would be trivial for the Word-only baseline.)

(a) Good examples where learned importance strengthens results according to the human evaluators.

(b) An easy case where all methods perform well, and close visual matches make Visual-only the best.

(c) A difficult case where no method can retrieve relevant images.

Fig. 15 Example results from the human evaluation experiments. The numbers on the right side of the images denote the fraction of people who selected the image as relevant to the query on the left. (**a**) For most retrieval results, we observe a clear gain in retrieval accuracy using our method, as judged by the human evaluators. (**b**) When there are near-exact matches, people are less likely to mark as relevant those retrieval results that contain the object from a correct category, but with different visual aspect. (**c**) If the query image is too difficult, then all three methods can fail

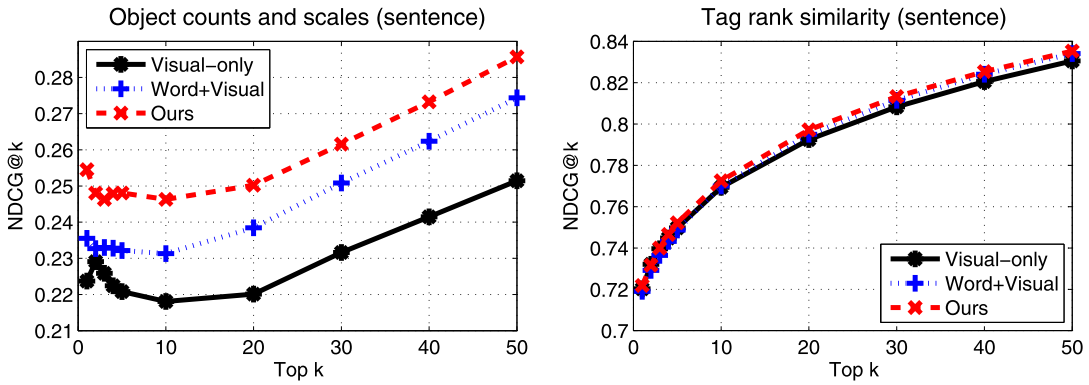


Fig. 16 Image-to-image retrieval results for the PASCAL+sentence dataset. Higher curves are better. See text for details

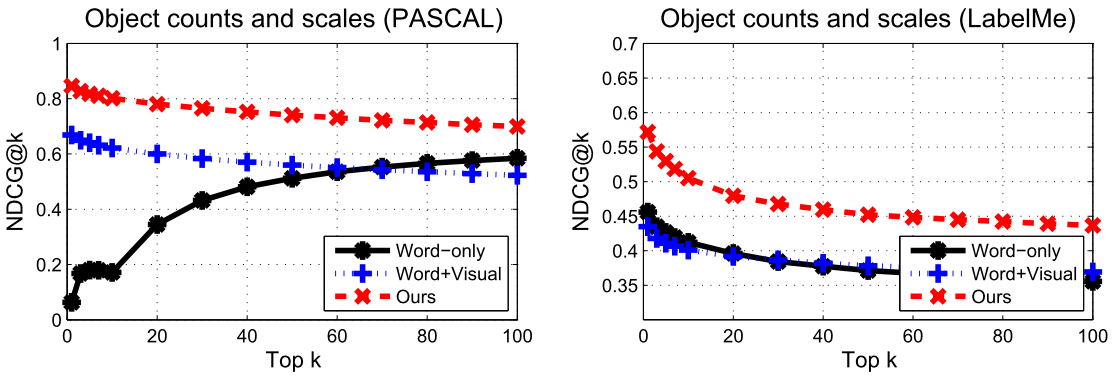


Fig. 17 Tag-to-image retrieval results. Given a user's ordered tag-list, our method retrieves images better respecting the objects' relative importance

Dataset	PASCAL VOC 2007+Tags				LabelMe+Tags			
Method	$K=1$	$K=3$	$K=5$	$K=10$	$K=1$	$K=3$	$K=5$	$K=10$
Visual-only	0.0826	0.1765	0.2022	0.2095	0.3940	0.4153	0.4297	0.4486
Word+Visual	0.0818	0.1712	0.1992	0.2097	0.3996	0.4242	0.4407	0.4534
Ours	**0.0901**	**0.1936**	**0.2230**	**0.2335**	**0.4053**	**0.4325**	**0.4481**	**0.4585**

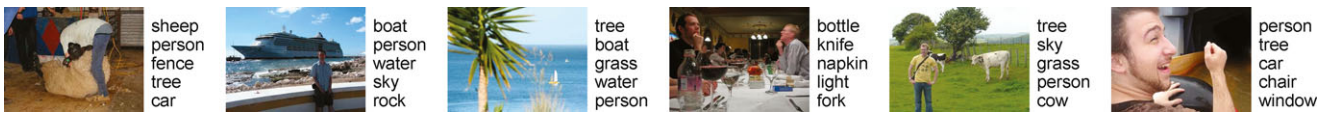


Fig. 18 *Top*: Image-to-tag auto annotation accuracy on the two tagged image datasets, as measured with the F1 score. Higher values are better. *Bottom*: Examples of auto-tagged images. Tags are quite accurate for images that depict typical scenes. Last example is a failure case

On PASCAL the Word-only baseline has very low accuracy on the top few retrieved results (see left plot in Fig. 17), likely because it cares only how many tags the images have in common with the query, whether they are foreground or background. That makes the baseline prone to retrieve images with the same background object tags (e.g., grass, sky, water) but possibly different foreground objects. This is penalized heavily by the object counts and scales reward function, which focuses on the foreground objects; on the more densely tagged LabelMe images, it is less of a pitfall for the baseline.

This result clearly illustrates the limitations of traditional keyword search retrieval: in terms of perceived importance, simply looking for images that share terms with a person's query words is insufficient. In contrast, by learning connections between how humans describe images and the appropriate relative layout of objects, we can obtain more consistent results.

We also note that the absolute magnitude of the NDCG scores are higher for this task than for content-based image retrieval; ours is almost double what it is for the object counts and scales reward function on the image-to-image results. We attribute this to the fact that keywords (tags) can serve as a more precise query, whereas a example-based image query may leave more ambiguity about the desired content.

4.7 Image-to-Tag Auto-Annotation Results

Finally, we explore auto-annotation, where our method takes an image and generates a list of tags. In contrast to previous work, however, we account for the importance of the tags when scoring the outputs: listing all the objects present is less valuable than listing those that most define the scene. We take the average ranks from the top K neighbors, and sort the keywords accordingly (see Sect. 3.3.3).

Figure 18 shows the results as a function of K, when using the two tagged image datasets, and Fig. 19 shows the results on the PASCAL when using the natural language captions to train and score results. We quantify the accuracy of the estimated output list using the F1 score, which accounts for both the precision and recall of the words included relative to the human-provided ground truth.

In these results, we observe that the Word+Visual approach, lacking any notion of importance, does not improve the auto-tagging accuracy over the Visual-only method. This is in concordance with the image-to-image retrieval results

Dataset	PASCAL VOC 2007+Sentences			
Method	$K = 1$	$K = 3$	$K = 5$	$K = 10$
Visual-only	0.0988	0.1313	0.1409	**0.1665**
Word+Visual	0.1057	0.1327	0.1399	**0.1592**
Ours	**0.1194**	**0.1435**	**0.1542**	0.1621

Fig. 19 Image-to-text auto annotation accuracy on the sentence data, as measured by the F1 score

measured using tag rank similarity in Fig. 9. When we consider only the top few retrieved images, the Visual-only method works about as well as it retrieves the near-exact matches that are likely to contain very similar tags, while the Word+Visual method retrieves images that have the same tags.

Overall, our method outperforms the baselines noticeably on the PASCAL images; differences are more modest on LabelMe, again likely due to the minor variation of importance per object occurrence. Given that PASCAL stems from real Flickr images, it is more realistic for the target setting where a user uploads photos and would like them auto-tagged and indexed. The fact that our results are strongest for this challenging set is therefore quite promising.

5 Conclusions

We proposed an unsupervised approach to learn the connections between human-provided tags and visual features, and showed the impact of accounting for importance in several retrieval and auto-tagging tasks. The key novelty is to reveal implied cues about object importance based on how people naturally annotate images with text, and then translate those cues into a dual-view semantic representation.

Our results show our method makes consistent improvements over pure content-based search as well as a method that also exploits tags, but disregards their implicit importance cues. We also show that our approach translates to learning an importance-aware semantic space with images that have natural language captions, and confirm through a human subject experiment that people find our method produces results with stronger perceived importance overall.

Through a series of experiments with multiple datasets, we have demonstrated the strengths and weaknesses of the proposed technique. On the whole, we see that for complex scenes in which objects play varying roles, our learned representation is much more suited to retrieval and auto-tagging. At the same time, we find that if using strong image descriptors (as we do here), a method that searches purely with the visual content is best when the database contains near-exact matches with similar scene layouts. This was evident in the LabelMe and human evaluation results. This means that for applications where the query is meant to bring up near matches (e.g., to find shots with a similar atmosphere, like the classic CBIR example of searching for other sunset images), using our learned representation is not necessary. More generally, these findings suggest that one might consider ways to automatically trade off the influence of close visual appearance matches with good contextual tag-based matches.

Our human subject tests show there is a fairly good level of agreement in practice about how well images relate in terms of important objects. This agrees with findings in previous work (von Ahn and Dabbish 2004; Spain and Perona 2008; Einhauser et al. 2008), and lends necessary support for our basic premise to help people find images with the "right" important objects. On the other hand, we also discovered that viewers are naturally influenced by the context in which image examples are displayed. Once near matches are visible, they are preferred over images that have similar objects but varying viewpoints or label granularities—at least without any further context about the target application. This nicely echoes our findings with the NDCG-scored results, as described above.

Our results using natural language data are promising, and they suggest that the simple rank-based cues can also play a role for learning with free-form descriptions. In future work, we plan to explore more elaborate feature extraction for natural language annotations.

Our approach aims to learn a single feature space capturing importance. While such a universal model of importance capturing what is salient to "any" observer is clearly appealing, it could also be interesting to consider user-specific models of importance. One could use our basic framework to build models for individual annotators, or groups of annotators instructed to generate their descriptions in the context of a particular *task* or application. For example, annotators tasked with understanding the interpersonal relationships of people captured in an image/video collection would give one form of descriptions, and annotators tasked with understanding the typical traffic flow would give another. Importantly, we would expect to see stronger agreement about the appropriate description among those interested in the same task, and can therefore tailor the learned representation most effectively.

In conclusion, this work shows how to account for the perceived importance of objects when performing content-based and cross-modal retrieval. Results on several data sets and textual sources indicate its advantages and potential pitfalls, and we think the analysis suggests several interesting directions for future work.

Acknowledgements We thank the anonymous reviewers for their constructive feedback and helpful suggestions to improve this manuscript. This research is supported in part by the Luce Foundation and DARPA CSSG N11AP20004.

References

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *CHI*.

Akaho, S. (2001). A kernel method for canonical correlation analysis. In *International meeting of Psychometric Society*.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Reading: Addison Wesley.

Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., & Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research*, *3*, 1107–1135.

Bekkerman, R., & Jeon, J. (2007). Multi-modal clustering for multimedia collections. In *CVPR*.

Berg, T., Berg, A., Edwards, J., & Forsyth, D. (2004). Who's in the picture. In *NIPS*.

Blaschko, M. B., & Lampert, C. H. (2008). Correlational spectral clustering. In *CVPR*.

Bruce, N., & Tsotsos, J. (2005). Saliency based on information maximization. In *NIPS*.

Datta, R., Joshi, D., Li, J., & Wang, J. (2008). Image retrieval: ideas, influences, and trends of the New Age. *ACM Computing Surveys*, *40*(2), 1–60.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In *CVPR*.

Duygulu, P., Barnard, K., de Freitas, N., & Forsyth, D. (2002). Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *ECCV*.

Einhauser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 1–26.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 1–15.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2007). The PASCAL visual object classes challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Farhadi, A., Hejrati, M., Sadeghi, A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: generating sentences for images. In *ECCV*.

Fergus, R., Fei-Fei, L., Perona, P., & Zisserman, A. (2005). Learning object categories from Google's image search. In *ICCV*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Fyfe, C., & Lai, P. (2001). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, *10*, 365–374.

Gupta, A., & Davis, L. (2008). Beyond nouns: exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*.

Hardoon, D., & Shawe-Taylor, J. (2003). KCCA for different level precision in content-based image retrieval. In *Third international workshop on content-based multimedia indexing*.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: an overview with application to learning methods. *Neural Computation*, *16*(12).

Hotelling, H. (1936). Relations between two sets of variants. *Biometrika*, *28*, 321–377.

Hwang, S. J., & Grauman, K. (2010a). Accounting for the relative importance of objects in image retrieval. In *British machine vision conference*.

Hwang, S. J., & Grauman, K. (2010b). Reading between the lines: object localization using implicit cues from image tags. In *CVPR*.

Jarvelin, K., & Kekalainen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, *45*(2), 83–105.

Kulis, B., & Grauman, K. (2009). Kernelized locality-sensitive hashing for scalable image search. In *ICCV*.

Lavrenko, V., Manmatha, R., & Jeon, J. (2003). A model for learning the semantics of pictures. In *NIPS*.

Li, L., Wang, G., & Fei-Fei, L. (2007). Optimol: automatic online picture collection via incremental model learning. In *CVPR*.

Li, L. J., Socher, R., & Fei-Fei, L. (2009). Towards total scene understanding: classification, annotation and segmentation in an automatic framework. In *CVPR*.

Li, Y., & Shawe-Taylor, J. (2006). Using KCCA for Japanese-English cross-language information retrieval and document classification. *Journal of Intelligent Information Systems*, *27*(2).

Loeff, N., & Farhadi, A. (2008). Scene discovery by matrix factorization. In *ECCV*.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, *60*(2).

Makadia, A., Pavlovic, V., & Kumar, S. (2008). A new baseline for image annotation. In *ECCV*.

Monay, F., & Gatica-Perez, D. (2003). On image auto-annotation with latent space models. In *ACM multimedia*.

Qi, G. J., Hua, X. S., & Zhang, H. J. (2009). Learning semantic distance from community-tagged media collection. In *ACM multimedia*.

Quack, T., Leibe, B., & Gool, L. V. (2008). World-scale mining of objects and events from community photo collections. In *CIVR*.

Quattoni, A., Collins, M., & Darrell, T. (2007). Learning visual representations using images with captions. In *CVPR*.

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2005). *Labelme: a database and web-based tool for image annotation* (Tech. rep). MIT.

Schroff, F., Criminisi, A., & Zisserman, A. (2007). Harvesting image databases from the web. In *ICCV*.

Smeulders, A., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1349–1380.

Spain, M., & Perona, P. (2008). Some objects are more equal than others: measuring and predicting importance. In *ECCV*.

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*, 643–659.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision*, *53*(2), 169–191.

Vijayanarasimhan, S., & Grauman, K. (2008). Keywords to visual categories: multiple-instance learning for weakly supervised object categorization. In *CVPR*.

Wolfe, J., & Horowitz, T. (2004). What attributes guide the deployment of visual attention and how do they do it? *Neuroscience*, *5*, 495–501.

Yakhnenko, O., & Honavar, V. (2009). Multiple label prediction for image annotation with multiple kernel correlation models. In *Workshop on visual context learning, in conjunction with CVPR*.