Accounting for the Relative Importance of Objects in Image Retrieval

Sung Ju Hwang
sjhwang@cs.utexas.edu
Kristen Grauman
grauman@cs.utexas.edu
The University of Texas
Austin, TX, USA

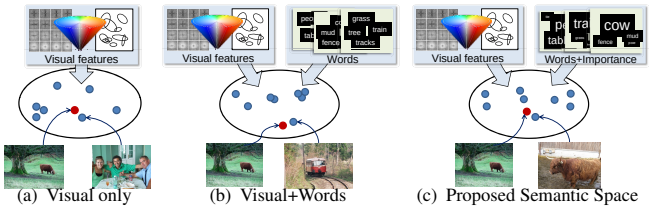

Figure 1: (a) Using only visual cues. (b) Learning a Visual+Words "semantic space". (c) Our idea: Learn the semantic space also using the order and relative ranks of the human-provided tag-lists.

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 and various labeling projects in the vision community. Often the keywords reflect the objects and events of significance and can thus be exploited as a loose form of labels and context. Researchers have explored a variety ways to leverage images with associated texts, including learning the correspondence between them for auto-annotation of regions, objects, and scenes, and building richer image representations based on the two simultaneous "views" for retrieval.

Existing approaches largely assume that image tags' value is purely in indicating the presence of certain objects. However, this 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. For example, if a system were to auto-tag the bottom right image in Figure 1(c) with *either* 'mud' or 'fence' or 'pole' or 'cow', not all responses are equally useful. Arguably, it is more critical to name those objects that appear more prominent or best define the scene (say, 'cow' in this example). Likewise, in image retrieval, the system should prefer to retrieve images that are similar not only in terms of their total object composition, but also in terms of those objects' relative importance to the scene.

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 treating tags simply as a set of names, we consider them as an ordered list. Specifically, we record a tag-list's nouns, their absolute ordering, and their relative rank compared to their typical placement. We propose an unsupervised approach based on Kernel Canonical Correlation Analysis (KCCA) 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 queries.

The three tag cues are defined as follows:

Word Frequency is a traditional bag-of-words that records the presence and count of each object. Each tag-list is mapped to an V-dimensional vector $W = [w_1, \ldots, w_V]$, where w_i is the number of times the i-th word is mentioned, and V is the vocabulary size. This feature serves to help learn the connection between the low-level image features and the objects they refer to.

Relative Tag Rank encodes the relative rank of each word compared to its typical rank: $R = [r_1, \ldots, r_V]$, where r_i is the percentile of the i-th word's rank relative to all its previous ranks observed in the training data. This feature captures the order of mention, which hints at the relative importance.

Absolute Tag Rank encodes the absolute rank of each word: $A = [\frac{1}{\log_2(1+a_1)}, \ldots, \frac{1}{\log_2(1+a_V)}]$, where a_i is the average absolute rank of the i-th word in the tag-list. In contrast to the relative rank, this feature more directly captures the importance of each object in the *same* scene.

For the image features, we use a diverse set of standard descriptors: Gist, color histograms, and bag-of-visual-words (BOW) histograms.

To leverage the extracted features to improve image retrieval, we use KCCA to construct a common representation (or semantic space) for both

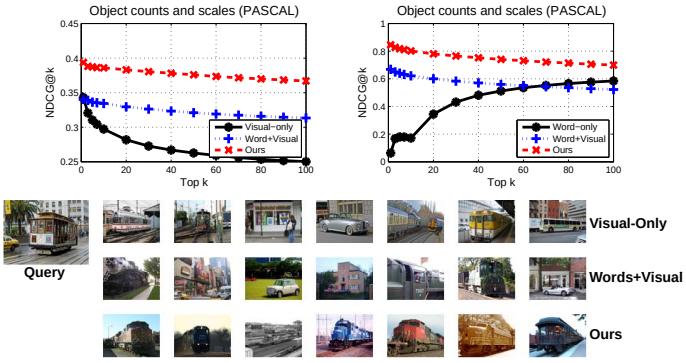

Figure 2: **Top**: Image-to-image and Tag-to-image retrieval results for PASCAL. **Bottom**: Example image-to-image retrieval. By modeling implied cues from the tag-lists when building the semantic space, our method retrieves images that better respect the objects' relative importance.

views of the data. KCCA is a kernelized version of CCA, which uses data consisting of paired views to simultaneously find projections from each feature space such that the correlation between projected features originating from the same instance is maximized. Given kernel functions for both feature spaces, $k_x(x_i, x_j) = \phi_x(x_i)^T \phi_x(x_j)$ and $k_y(y_i, y_j) = \phi_y(y_i)^T \phi_y(y_j)$, KCCA seeks projection vectors in the kernels' implicit feature spaces, w_x and w_y, where $w_x = \sum_i \alpha_i \phi_x(x_i)$ and $w_y = \sum_i \beta_i \phi_y(y_i)$. The objective is to identify the $\alpha, \beta \in \Re^N$ that maximize

$$\max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \beta^T K_y^2 \beta}}, \quad (1)$$

where K_x and K_y are $N \times N$ kernel matrices over a sample of N pairs. The optimization can be formulated as an eigenvalue problem (see Hardoon et al.,2004).

In our case, $k_x(x_i, x_j)$ is an average of kernels for the visual features, while $k_y(y_i, y_j)$ averages the tag feature kernels. After learning the semantic space, we project all database images onto it. Then we can perform three different tasks:

- **Image-to-Image Retrieval:** Given a novel query image, we use its image features only to project onto the semantic space, and then sort all database images relative to it based on their normalized correlation in that space. The results should be favorably skewed towards showing scenes with similarly relevant objects when using our approach.

- **Tag-to-Image Retrieval:** Given a novel tag-list query, we project onto the semantic space, and again sort the database images by correlation. We expect to retrieve images where not only are objects shared with the query, but they are also of similar relative importance.

- **Image-to-Tag Auto-Annotation:** Given a novel query image, we project onto the semantic space, identify its nearest example among those that are tagged, and return the top keywords on that tag-list. This strategy attempts to provide the most *important* tags based on all available image features.

The results show that our method makes consistent improvements over a pure visual search baseline, as well as a method that also exploits tags, but disregards their implicit importance cues (Fig. 2).

In short, our main contributions are: (1) an approach to learn the relative importance 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 quantitative measures of object prominence.