Keywords to Visual Categories: Multiple-Instance Learning for Weakly Supervised Object Categorization

Sudheendra Vijayanarasimhan and Kristen Grauman

University of Texas at Austin
1 University Station C0500
Taylor Hall 2.124
University of Texas at Austin
Austin, TX 78712-1188

{svnaras, grauman}@cs.utexas.edu
http://www.cs.utexas.edu/~grauman/

Conventional supervised methods for image categorization rely on manually annotated (labeled) examples to learn good object models, which means their generality and scalability depends heavily on the amount of human effort available to help train them. The Web is an alluring source of image data for vision researchers, given both the scale at which images are freely available as well as the textual cues that surround them. Querying a keyword-based search engine (e.g., Google Image Search) or crawling for meta-tags (e.g., on Flickr) will naturally yield images of varying degrees of relevance: only a portion will contain the intended category at all, others may contain instances of its homonym, and in others the object may barely be visible due to clutter, low resolution, or strong viewpoint variations.

Though appealing, it is of course difficult to learn visual category models straight from the automatically collected image data. Recent methods attempt to deal with the images' lack of homogeneity indirectly, either by using clustering techniques to establish a mixture of possible visual themes [5, 2, 3], or by applying models known to work well with correctly labeled data to see how well they stretch to accommodate "noisily" labeled data [1, 4]. Unfortunately, the variable quality of the search returns and the difficulty in automatically estimating the appropriate number of theme modes make such indirect strategies somewhat incompatible with the task.

In this work, we propose a more direct approach to learn discriminative category models from images associated with keywords. We show that multiple-instance learning enables the recovery of robust category models from images returned by keyword-based search engines. Given a list of category names, our method gathers groups of potential images of each category via a number of keyword-based searches on the Web. Because the occurrence of true exemplars of each category may be quite sparse, we treat the returned groups as positive bags that contain some unknown amount of positive examples, in addition to some unrelated negative examples. Complementary negative bags are obtained by collecting sets of images from unrelated queries, or alternatively from any existing database having categories outside of the input list.

We show how optimizing a large-margin objective function with constraints that reflect the expected sparsity of true positive examples yields discriminative models that can accurately predict the presence of the object categories within novel images, and/or provide a good re-ranking of the initial search returns. In addition, we show how to iteratively improve the learned classifier by automatically refining the representation of the ambiguously labeled examples.

Our learning paradigm exploits the wealth of text surrounding natural images that already exists, while properly accounting for their anticipated noise and ambiguity. Experimental results indicate the approach's promise: on benchmark image datasets it competes well with several fully supervised methods, is more accurate than a single-instance learning SVM baseline, and improves on state-of-the-art unsupervised image classification results.

References

[1] R. Fergus, P. Perona, and A. Zisserman. A Visual Category Filter for Google Images. In *Proceedings of the European Conference on Computer Vision* (ECCV), Prague, Czech Republic, May 2004.

[2] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google's Image Search. In *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), Beijing, China, October 2005.

[3] L. Li, G. Wang, and L. Fei-Fei. OPTIMOL: Automatic Online Picture Collection Via Incremental Model Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Minneapolis, MN, June 2007.

[4] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting Image Databases from the Web. In *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), Rio de Janeiro, Brazil, October 2007.

[5] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering Object Categories in Image Collections. In *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), Beijing, China, October 2005.

Topic: visual processing and pattern recognition
Preference: poster or oral