# Exemplar-SVMs for Visual Object Detection, Label Transfer and Image Retrieval

**Tomasz Malisiewicz** TOMASZ@CSAIL.MIT.EDU
Massachusetts Institute of Technology

**Abhinav Shrivastava** ASHRIVAS@CS.CMU.EDU
**Abhinav Gupta** ABHINAVG@CS.CMU.EDU
**Alexei A. Efros** EFROS@CS.CMU.EDU
Carnegie Mellon University

Today's state-of-the-art visual object detection systems are based on three key components: 1) sophisticated features (to encode various visual invariances), 2) a powerful classifier (to build a discriminative object class model), and 3) lots of data (to use in large-scale hard-negative mining). While conventional wisdom tends to attribute the success of such methods to the ability of the classifier to generalize across the positive class instances, here we report on empirical findings suggesting that this might not necessarily be the case. We have experimented with a very simple idea: to learn a separate classifier *for each positive object instance* in the dataset (see Figure 1). In this setup, no generalization across the positive instances is possible by definition, and yet, surprisingly, we did not observe any drastic drop in performance compared to the standard, category-based approaches.

More specifically, we train a separate linear SVM for every exemplar in the training set (e.g., every annotated bounding box in case of object detection). Each of these Exemplar-SVMs is thus defined by a single positive instance and millions of negatives. Taken together, an Ensemble of Exemplar-SVMs (Malisiewicz et al., 2011), aims to combine the effectiveness of a discriminative object detector with the explicit correspondence offered by a nearest-neighbor approach. While each detector is quite specific to its exemplar, we empirically observe that, after a simple calibration step, an ensemble of such Exemplar-SVMs offers surprisingly good performance, roughly comparable to the much more complex latent part-based model of (Felzenszwalb et al., 2010).

It is interesting to note some of the properties of Exemplar-SVMs:

- There is little sign of overfitting. Although each SVM has only a single positive example, the huge number of negatives appear to provide enough to constrain the problem. In a way, the exemplar's decision boundary is defined, in large part, by what it is *not*. At the same time, each classifier is solving a much simpler problem than in the full category case.
- While the large imbalance between the positive and negative sets can often lead to a poor decision boundary, we have empirically found that the induced ordering of the detections with respect to that boundary is still good. Thus, the Exemplar-SVM can be interpreted as ordering the negatives by visual similarity to the exemplar.
- Exemplar-SVMs are related to learning per-exemplar distance functions (Frome & Malik, 2006; Malisiewicz & Efros, 2008). The crucial difference between a per-exemplar classifier and a per-exemplar distance function is that the latter forces the exemplar itself to have the maximally attainable similarity. An Exemplar-SVM has much more freedom in defining the decision boundary and is better able to incorporate input from the negative samples.
- While a standard category classifier treats positive and negative examples in the same way, the Ensemble of Exemplar-SVMs handles them differently. One way to think about it is that the positives are represented non-parametrically, while the negatives are represented parametrically.

In addition to being an interesting empirical result, there are a number of potential advantages of the Ensemble of Exemplar-SVMs formulation compared to standard, category-based methods:

- Detections show good alignment with the corresponding source exemplar, making it possible to transfer any available exemplar meta-data (segmentation, geometric structure, 3D model, etc) directly onto the detection (see Figure 2).
- Since learning is exemplar-specific, there is no need to map all exemplars into a common feature space. Therefore individual exemplars can be represented in

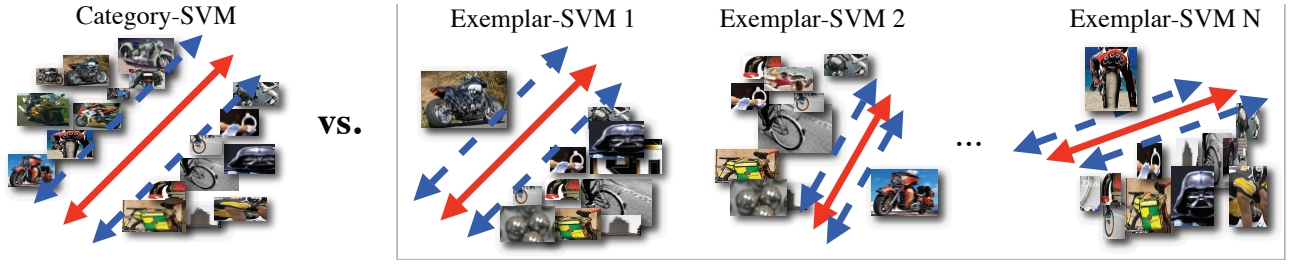


*Figure 1.* **An Ensemble of Exemplar-SVMs (Malisiewicz et al., 2011).** Instead of placing all positives into a single category-specific learning problem, we train a separate linear SVM for each positive instance in the dataset. One way to think about it is that the positives are represented non-parametrically, while the negatives are represented parametrically.

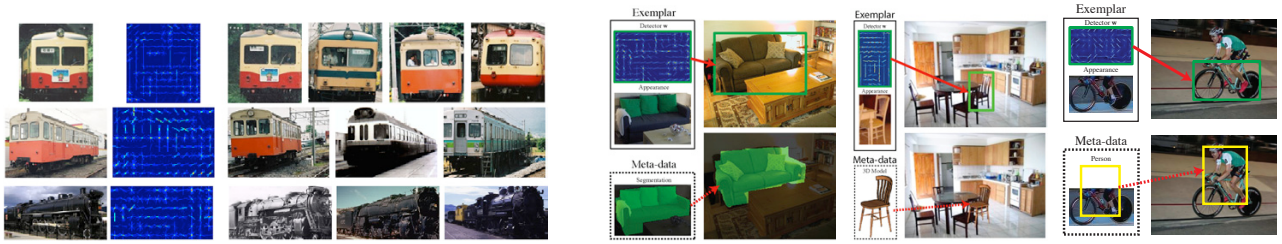


*Figure 2.* **Detection and Transfer via Exemplar-SVMs.** Left: A few "train" exemplars with their top detections on the PASCAL VOC test-set. Right: We can transfer meta-data such as segmentation, exemplar-aligned 3D model, or extra annotations (e.g., person) directly onto the detection window.


*Figure 3.* **Large-scale Image Matching (Shrivastava et al., 2011).** Two image matching results showing the top three images retrieved by Exemplar-SVMs as well as the top three images returned by Google's Visual Image Search.

different feature spaces (i.e., different template sizes as well as entirely different features).

- Because of the long-tailed distribution of objects in the world (10% of objects own 90% of exemplars), the extra cost of using exemplars vs. categories will greatly diminish as the number of categories increases. Moreover, learning is embarrassingly parallel – instead of solving a single large and non-convex optimization problem (Felzenszwalb et al., 2010), here each Exemplar-SVM's objective function is convex and can be optimized independently.

Additionally, we have observed that even when the negative set is not completely clean (i.e., happens to contain some instances from the positive class), this has only a modest detrimental effect on object detection performance (3.2% drop on the PASCAL VOC 2007 (Malisiewicz, 2011)). Such robustness to the negative set being polluted by in-class instances motivated us to experiment with using the Exemplar-SVM formulation for the task of **large-scale image retrieval** (Shrivastava et al., 2011). Here, the query image is treated as the single positive exemplar and a collection of tens of thousands random unlabeled Flickr images serves as the negative set. The resulting SVM weight vector defines a new distance which can be used to retrieve images for this particular query (see Figure 3). We demonstrate the usefulness of this approach with results on matching images across different visual domains, such as photos taken over different lighting-conditions, paintings, and sketches.

## References

Felzenszwalb, P., Girschick, R., McCallester, D., and Ramanan, D. Object detection with discriminatively trained part based models. *PAMI*, 2010.

Frome, A. and Malik, J. Image retrieval and recognition using local distance functions. *NIPS*, 2006.

Malisiewicz, T. Exemplar-based representations for object detection, association and beyond. *Carnegie Mellon University PhD Thesis*, 2011.

Malisiewicz, T. and Efros, A. A. Recognition by association via learning per-exemplar distances. *CVPR*, 2008.

Malisiewicz, T., Gupta, A., and Efros, A. A. Ensemble of exemplar-svms for object detection and beyond. *ICCV*, 2011.

Shrivastava, A., Malisiewicz, T., Gupta, A., and Efros, A. A. Data-driven visual similarity for cross-domain image matching. *SIGGRAPH Asia*, 2011.