Discriminatively Trained Mixtures of Deformable Part Models

Pedro Felzenszwalb and Ross Girshick
University of Chicago
pff,rbg@cs.uchicago.edu

David McAllester
TTI-C
mcallester@tti-c.org

Deva Ramanan
UC Irvine
ramanan@tti-c.org

We have developed a new system building on our work on discriminatively trained, multiscale, deformable part models [1]. As in our previous system the models are trained using a discriminative procedure that only requires bounding box labels for positive examples.

Our new system uses mixture models. Each mixture component is similar to a model from [1], consisting of a coarse "root" filter and 6 part models. Each part model consists of a spatial term and a part filter. The spatial term specifies an ideal location for a part relative to the root and a quadratic deformation cost for placing the part at some other location. The score of a component in a detection window is the score of the root filter on the window plus the sum over parts, of the maximum over placements of that part, of the part filter score on its subwindow minus the deformation cost of the placement. Both root and part filters are scored by computing the dot product between a set of weights and histogram of gradient (HOG) features within a window. As in [1] the features for the part filters are computed at twice the spatial resolution of the root filter. The score of a mixture model in a detection window is the maximum score over its components, where the scores are calibrated by a component specific offset parameter. Models are defined at a fixed scale, and we detect large objects by searching over an image pyramid.

The main innovations of the new system are:

1. Each object model now consists of a mixture of two components, where each component is itself a multiscale deformable part model. This makes it possible to better capture different aspects of an object. For example, for cars one component can capture side views while the other component captures front and rear views.

2. We optimize the true latent SVM objective function using stochastic gradient descent together with a mechanism for data-mining hard negative examples. Our previous system used SVMlight for optimizing an approximation of the objective function. We have found that optimizing the correct objective function is important when training mixture models.

3. We use new HOG features that are both lower-dimensional and sensitive to contrast direction. Reducing the dimensionality speeds up the detector while the contrast sensitivity improves its accuracy on a number of classes.

4. The search over scales is done at a coarser resolution. This speeds up the detector at the expense of a small reduction in localization accuracy. However, we now predict an object bounding box by using a linear function of the location of its parts. This improves the localization accuracy for some classes.

5. We use a simple procedure to rescore each detection using a "context" descriptor that is defined in terms of the maximum score of each object class in an image.

Model Initialization To initialize the mixture models we split the positive examples in each object category into two groups, according to the aspect ratio of their bounding boxes. We initialize each mixture component by training root filters and selecting parts from those filters using a procedure similar to [1].

Model Update After initialization each mixture model is trained using the latent SVM framework, where we alternate between estimating latent parameters for the positive examples and optimizing the latent SVM objective function. In the case of mixture models the latent parameters in each example include a component label and the placement of the filters associated with the corresponding mixture component.

When the latent parameters for positive examples are fixed the latent SVM objective function is convex. We use stochastic gradient descent to optimize the objective, together with a caching mechanism similar to the one in [1] for data-mining hard negative examples.

HOG features The system in [1] used 36-dimensional HOG features, capturing 9 gradient orientations under 4 normalizations. We have found that we can use 13-dimensional features, capturing 9 orientations under a single normalization plus 4 features capturing texture gradients, without reducing detection accuracy. The reduced feature set can be seen as a linear projection of the original features which preserves most of their information. An expansion to 18 contrast sensitive orientations and 4 texture gradients leads to improved performance on some classes. For the final classifiers we used 31-dimensional HOG features that include 18 contrast sensitive and 9 contrast insensitive orientations, and 4 texture gradients.

Context Rescoring We use a simple procedure to rescore detections using contextual information. We define the context of an image by a 20-dimensional vector specifying the maximum score of each object detector in the image (one dimension per class). To rescore a detection we build a 25-dimensional feature vector containing: the original score of the detection, the top-left and bottom-right bounding box coordinates, and the image context. We use a class-specific SVM to classify this feature vector, leading to a new score for the detection. This SVM is trained from labeled data generated from the "train+val" datasets. We simply run our detectors on the training data and label each detection as true positive or false positives using the PASCAL annotations. This rescoring procedure, though very simple, leads to a small improvement in average precision on several classes in the 2007 dataset.

References

[1] P. Felzenszwalb, D. McAllester, and D. Ramaman. A discriminatively trained, multiscale, deformable part model. In *IEEE CVPR*, 2008.