The Pascal Visual Object Classes Challenge 2008 submission

Adrien Gaidon, Marcin Marszałek, Cordelia Schmid

September 29, 2008

1 Introduction

We submit two recognition methods based on the same underlying image representations that we call *channels*. Each channel is defined by a choice of an image sampler, a local descriptor and a global spatial grid. Given a channel and a visual vocabulary [8], a histogram of features can be used to represent each image and a similarity between images can be estimated. The submitted methods also share the classifier, which is a one-vs.-all non-linear Support Vector Machine [6].

The methods differ in how they combine multiple channels. The first method is based on the approach of Zhang et al. [9], where the distances corresponding to each channel are added to form a final image similarity measure. The second method employs an optimization algorithm, which is used to determine (on a per-class basis) the parameters of the generalized RBF kernel incorporating all the channels.

Note that our submission builds upon and extends LEAR's last year's submission [4] by improving the image representation as well as the optimization strategy.

2 Channels

We densely sample the images using a multi-scale grid and use Harris-Laplace [5], Hessian, Harris-Harris and Laplacian [2] interest points detectors. We mainly use the SIFT [3] to describe local regions of interest. From our own experiments and the results reported in [7], we also chose to use the Opponent-SIFT color descriptor.

We count the local features in 1x1, 2x2 and horizontal 3x1 grids covering the whole image [1]. Note that the 1x1 grid results in a standard bag-of-features representation.

This makes a total of 30 channels combined together.

3 Classifier

The classification is performed with a non-linear Support Vector Machine [6]. The multi-class problem is addressed in a one-vs.-all set-up. We have used the χ^2 distance to measure the similarity between images. We use two different kernels.

3.1 Flat approach

For the first submission we follow the kernel design of Zhang et al. [9]. We also fix the C parameter to the value suggested in the paper.

3.2 Shotgun approach

For the second submission we learn the generalized RBF kernels (one for each class) using a random-restart hill climbing algorithm (also called shotgun hill climbing). This allows us to learn the optimal product kernel stemming from the single channel kernels. It amounts to learn the importance of each sampling/description/spatial method for the recognition of each class separately.

References

[1] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[2] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2), 1998.

[3] D. Lowe. Distinctive image features form scale-invariant keypoints. *IJCV*, 60(2), 2004.

[4] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition, oct 2007. Visual Recognition Challange workshop, in conjunction with ICCV.

[5] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1), 2004.

[6] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA, 2002.

[7] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek. Evaluation of color descriptors for object and scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska*, 2008.

[8] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *IWLAVS*, 2004.

[9] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 2007.