Object Localization with Global and Local Context Kernels

Matthew B. Blaschko
http://www.robots.ox.ac.uk/~blaschko/
Christoph H. Lampert
http://www.kyb.mpg.de/~chl

Department of Engineering Science
University of Oxford, UK
Max Planck Institute for Biological Cybernetics
Tübingen, Germany

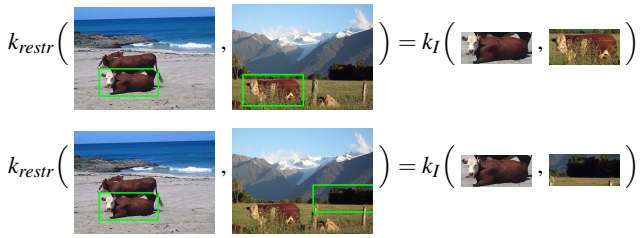
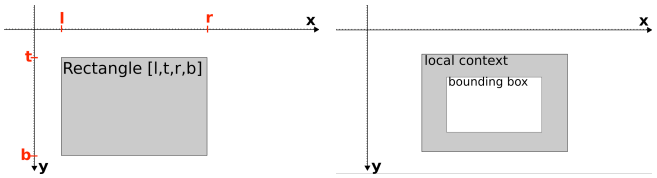


Figure 1: Illustration of the *restriction kernel*: (image, box) pairs are compared by restricting the image to box region and applying a traditional image kernel k_I to the resulting subimages. The kernel value in the top row will be larger than the one in the bottom row, because the subregions are more similar.



(a) Parameterization of a bounding box by its *left*, *top*, *right*, and *bottom* coordinates in the image plane.

(b) The spatial extent of a local context kernel is indicated by the shaded region.

Figure 2: The parameterization of a bounding box (a), and the spatial extent of a local context kernel (b).

We formalize our notions of global and local context in the framework of kernel classifiers, e.g. support vector machines [4]. In particular, we make use of the concept of *joint kernels* [1], positive definite functions operating jointly on both input and output spaces, denoted $\mathcal{X}$ and $\mathcal{Y}$, respectively. A joint kernel, $k : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, takes two input-output pairs as arguments, and returns a value of similarity between these pairs. As an example, we take the *restriction* kernel introduced in [2]. We denote an image in the space of possible images as $x \in \mathcal{X}$, and a bounding box as $y \in \mathcal{Y}$. The restriction kernel is defined by restricting (cropping) each image to its corresponding bounding box (denoted $x|_y$), and then applying a standard image kernel to the resulting cropped regions (Figure 1)

$$k_{restr}((x_i,y_i),(x_j,y_j)) = k_I(x_i|_{y_i}, x_j|_{y_j}). \tag{1}$$

Analogously, we define a *local context kernel* to be an image kernel on a region *around the object of interest*. In order to be invariant to scale change, we define the spatial extent of the contextual regions *relative* to the bounding box of interest. Specifically, if (l,t,r,b) defines the coordinates of the left, top, right, and bottom of a bounding box around the object of interest, respectively (Figure 2(a)), we define the contextual region to be the region *between* the box (l,t,r,b) and the larger rectangle $(l-\theta w, t-\theta h, r+\theta w, b+\theta h)$ where $w = r-l$ and $h = b-t$ are width and height of the bounding box (Figure 2(b)). The scalar θ parameterizes the size of the contextual region relative to the size of the bounding box. For $\theta = 1/\sqrt{2}$, the contextual region has the same area as the bounding box. We have used this value in the experiments. Using the notation $\Theta(y)$ to denote the contextual region defined by the parameter θ for a bounding box y, we define a context kernel analogously to the restriction kernel by also leveraging an existing image kernel, k_I,

$$k_{local}((x_i,y_i),(x_j,y_j);\theta) = k_I(x_i|_{\Theta(y_i)}, x_j|_{\Theta(y_j)}) \tag{2}$$

where we have made the kernel's dependence on θ explicit.

We define a *global context kernel* to be one that incorporates information from the entire image, but does not depend on the bounding box:

$$k_{global}((x_i,y_i),(x_j,y_j)) = k_I(x_i,x_j) \tag{3}$$

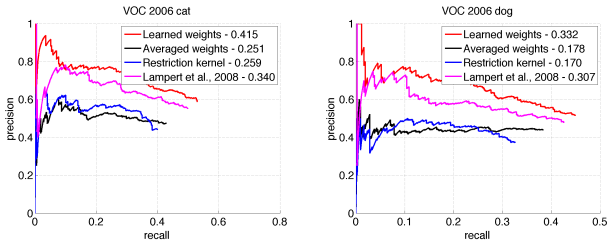


Figure 3: Precision-Recall curves for categories *cat* and *dog* of PASCAL VOC 2006.

for an arbitrarily chosen image kernel k_I. Given these ingredients, we define a joint kernel function that can perform object localization with global and local context

$$\begin{aligned} k((x_i,y_i),(x_j,y_j)) = \beta_1 k_{restr}((x_i,y_i),(x_j,y_j)) + \\ \beta_2 k_{local}((x_i,y_i),(x_j,y_j)) + \beta_3 k_{global}((x_i,y_i),(x_j,y_j)). \end{aligned} \tag{4}$$

Through the weight parameters $\beta_j > 0$ we can control the relative importance of the individual contributions. These parameters are learned using multiple kernel learning.

Localization is performed by maximizing the objective function with respect to the bounding box, y. As in [3], we use a branch and bound optimization, efficient subwindow search (ESS), to perform this maximization. ESS searches the space of possible bounding boxes by keeping a priority queue that stores *sets* of possible bounding boxes ordered by an upper bound, $\hat{f}$, on a given quality function, f.

Following [3] we represent sets of boxes Y as intervals over the left, top, right, and bottom coordinates of the bounding box in the image plane. Using ideas from *interval arithmetic*, we propagate the uncertainty in Y through the transformation Θ: denoting $Y = \left([\underline{l},\overline{l}],[\underline{t},\overline{t}],[\underline{r},\overline{r}],[\underline{b},\overline{b}]\right)$, we specify intervals

$$\begin{aligned} \overline{\Theta}(Y) = \Big(& [\underline{l}-\theta\overline{w}, \overline{l}-\theta\underline{w}], [\underline{t}-\theta\underline{w}, \overline{t}-\theta\underline{h}], \\ & [\underline{r}+\theta\underline{w}, \overline{r}+\theta\overline{w}], [\underline{b}+\theta\underline{h}, \overline{b}+\theta\underline{h}] \Big), \end{aligned} \tag{5}$$

where $\overline{w} = \overline{r}-\underline{l}$, $\overline{h} = \overline{b}-\underline{t}$, $\underline{w} = \max(0, \underline{r}-\overline{l})$ and $\underline{h} = \max(0, \underline{b}-\overline{t})$, such that $\overline{\Theta}(Y)$ specifies the intervals for the external boundary of the local contextual region parameterized by θ. We can write a valid upper bound compactly as:

$$\hat{f}_{local}(x,Y;\theta) = \overline{\Theta}(Y)^+_{max} + \overline{\Theta}(Y)^-_{min} - Y^+_{min} - Y^-_{max}. \tag{6}$$

We have performed experiments using the restriction kernel only, a flat weighting over all context kernels, and learned weights from MKL. Figure 3 shows the precision recall curves for the VOC 2006 categories *cat* and *dog* that were also reported in [3]. Learned weights with MKL consistently improve performance over the flat weighting and over the restriction kernel.

[1] G. H. Bakır, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan. *Predicting Structured Data*. The MIT Press, 2007.

[2] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In *ECCV*, 2008.

[3] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *CVPR*, 2008.

[4] B. Schölkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, 2002.