Object Localization with Global and Local Context Kernels

Matthew B. Blaschko
http://www.robots.ox.ac.uk/~blaschko/
Visual Geometry Group
Department of Engineering Science
University of Oxford, UK

Christoph H. Lampert
http://www.kyb.mpg.de/~chl
Max Planck Institute for Biological
Cybernetics, Tübingen, Germany

Abstract

Recent research has shown that the use of contextual cues significantly improves performance in sliding window type localization systems. In this work, we propose a method that incorporates both global and local context information through appropriately defined kernel functions. In particular, we make use of a weighted combination of kernels defined over local spatial regions, as well as a global context kernel. The relative importance of the context contributions is learned automatically, and the resulting discriminant function is of a form such that localization at test time can be solved efficiently using a branch and bound optimization scheme. By specifying context directly with a kernel learning approach, we achieve high localization accuracy with a simple and efficient representation. This is in contrast to other systems that incorporate context for which expensive inference needs to be done at test time. We show experimentally on the PASCAL VOC datasets that the inclusion of context can significantly improve localization performance, provided the relative contributions of context cues are learned appropriately.

1 Introduction

Sliding window classifiers in their original form attempt to decide the presence or absence of an object at a specific location using only local information. However, experiments in human psychophysics, *e.g.* by Palmer [27], indicate that context is a crucial cue in object detection. Torralba [34] and Bar [3] show that humans are capable of correctly identifying faces and objects of very similar or even identical appearance if they occur in their natural context. Consequently, the incorporation of contextual cues in computer vision tasks has received a large amount of attention.

There are many different forms of contextual cues that can be used. In the case of still image classification, recent work has exploited *external context*, such as EXIF tags [37], Flickr tags [30], and geo tags [25]. Rather than relying on meta-data, we focus in this work on the use of visual contextual cues that are present within the image itself. Visual context has been studied on different levels. As *global context*, Torralba et al. [35] and Murphy et al. [26] proposed to represent the full image by its *gist*, and to include this global representation as an additional feature in object classifiers Hoiem et al. [16] propose to first infer the 3D scene geometry from an image in order to help a subsequent object detection step.

BMVC 2009 doi:10.5244/C.23.63

On the other end of the size scale, context has been used on a per-pixel level. Introducing the concept of *Things and Stuff*, Adelson [1] shows that context is beneficial to identify *material* from otherwise ambiguous local features. Similarly, Shotton et al. [32] use the relative neighborhoods of pixels to improve object-based segmentation results, an approach that recently has been extended by Gould et al. [14] to also include the locations of pixels within a neighborhood. Lazebnik and Raginsky use pixel neighborhoods in a largely unsupervised fashion to improve per pixel classification [24].

Several approaches for the integration of higher level context rely on a pre-segmentation of the image. Information about the labels of neighboring then allows better classification of each segment. Corresponding models have been proposed *e.g*. Baumgartner et al. [4] for road detection in aerial images, Singhal et al. [33] for scene classification and [8] for image labeling. Kumar and Hebert [18] extended this idea and constructed a two-layer Markov Random field that is able to balance the pixel-level evidence against the context information from the neighboring image segments' class labels. Rabinovich et al. [28] also use the relation between neighboring segments, but they propose a two-stage procedure that first classifies each region separately, and then performs a post-processing operation that can change the region labels based on the observed context. Similar post-processing operations have also been used in face detection, *e.g*. by Bergboer et al. [6].

Our own work targets object localization in images, a topic for which context has also successfully been applied: Kruppa and Schiele use a fixed region surrounding a detection window to improve the detection of face with very low resolution [17]. Dalal and Triggs showed that one achieves better results in pedestrian detection if a detection window larger than the actual person is used [9]. Uijlings et al. evaluate the best size of bounding box to improve localization performance [38]. An empirical evaluation of several different kinds of contextual cues is given in [10].

Although the dominant opinion is that the inclusion of context is always helpful, Wolf and Bileschi [39] argue against this view and show that the relevance of context depends strongly on the situation at hand. Therefore, it makes sense to not use fixed context models, but to learn also about the context from data. This has specifically been proposed in multi-class boosting scenario, *e.g*. by Fink and Perona [13] and Torralba et al. [36]. Heitz and Koller [15] construct a probabilistic model that learns which *stuff* in an image helps in the identification of *things*. For multi-class object localization, Lampert and Blaschko [20] propose to learn a discriminative classifier that takes into account which object class is useful as context for which other class, and apply it as post-processing operation to the detections of a context-unaware detection system.

The primary contributions of this work is two-fold. Firstly, we introduce the concept of *global and local context kernels* that allow us to combine different context models into a single discriminative kernel classifiers, learning the importance of each contributions as part of the training step. Secondly, we show how to integrate the resulting context-aware kernels into the recently proposed *efficient subwindow search* framework for object localization [21, 22], thereby allowing extremely efficient evaluation.

2 Global and Local Context Kernels

We formalize our notions of global and local context in the framework of kernel classifiers, *i.e*. support vector machines [31]. In particular, we make use of the concept of *joint kernels* [2]. Joint kernels are positive definite functions operating jointly on both input and

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.

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 a concrete example, we take the *restriction* kernel introduced in [7]. 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)

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

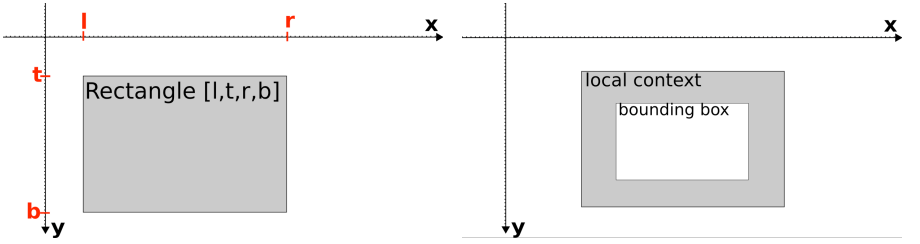
Analogously, we define a *local context kernel* to be an image kernel on a region *around the object of interest*. The spatial extent of this region defines the amount of local context to use. In order to be invariant to scale change, we define the spatial extent of the contextual regions *relative* 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 in Section 5. 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.[1]

In contrast to local context, we define a *global context kernel* to be one that incorporates

[1] Note that because the region $\Theta(y)$ is not a rectangle, but the difference region between two rectangles, not all kernels defined for images might be applicable for the context region. However, most popular image kernels are able to handle regions of such shape, in particular the ones based on bag-of-visual-word histograms that we use for our experiments.



(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 as the left, top, right, and bottom in the image plane (a), and the spatial extent of a local context kernel (b).

information from the entire image, but does not depend on the bounding box coordinates:

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

for an arbitrarily chosen image kernel k_I. Note that this definition incorporates previous notions of global contextual including *gist* [26, 35].

In choosing the restriction kernel, context kernels, and global context kernels, we are free to rely on different image kernels. This is of particular interest in the context of branch-and-bound optimization (as will be presented in Section 4) where we require efficiently computable bounding functions for the restriction and local context kernels, while there is no such requirement for the global context kernel.

Given these ingredients, we define a joint kernel function that can perform object localization with global and local context

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

Through the weight parameters $\beta_j > 0$ we can control the relative importance of the individual contributions.

3 Learning Procedure

Let $\{(x_i,y_i)\}_{i=1,\dots,m}$ be a sample of training images and bounding boxes indicating the presence of an object of our current class. From the same images we sample additional bounding boxes that do not have significant overlap with the y_i, forming additional samples $\{(x_i,y_i)\}_{i=m+1,\dots,n}$. We combine both sets into a training set of images and bounding boxes, $\{(x_i,y_i,\ell_i)\}_{i=1,\dots,n}$, where $\ell_i = +1$ if y_i specifies the location of an instance of our object class, and $\ell_i = -1$ otherwise. From the representer theorem [31], it follows that the optimal maximum margin classifier (SVM) in this setup must have the form

$$f(x,y) = \sum_i \alpha_i k(x,y,x_i,y_i) + b \tag{5}$$

for some parameters α and b. Substituting the expression (4) for k, we obtain

$$f(x,y) = \sum_i \alpha_i \sum_j \beta_j k_j(x,y,x_i,y_i) + b \tag{6}$$

where the inner summation ranges over the *restriction*, *local*, and *global context* kernels. Because of the additive structure, it is possible to simultaneously learn optimal SVM parameters α and the weight coefficients β using *multiple kernel learning* [23, 29]. Once we have learned an appropriate function, f, that measures the quality of a bounding box y in an image x, we obtain a localization function, $g : \mathcal{X} \to \mathcal{Y}$, that predicts bounding boxes from images by selecting the best possible bounding box as measured by f:

$$g(x) = \operatorname{argmax}_y f(x,y). \tag{7}$$

In [21], an efficient branch-and-bound technique is proposed to solve (7) for classifiers based on the restriction kernel. In the next section we discuss to extend this approach in order to allow efficient object localization with the global and local context kernels.

4 Branch-and-Bound Optimization

Efficient subwindow search (ESS) is a branch and bound framework for object localization first introduced in [21]. 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. At each stage in the search procedure, the set of bounding boxes with highest upper bound is dequeued and split into two disjoint subsets. Each of these are inserted into the priority queue with an upper bound computed over a smaller region of uncertainty [21, 22]. When an item is dequeued that consists of only one bounding box, we have converged to the solution. The returned bounding box is optimal over *all* possible bounding boxes given two conditions on the upper bound, $\hat{f}$:

$$\hat{f}(x,Y) \geq f(x,y) \quad \text{for all } y \in Y, \tag{8}$$

$$\hat{f}(x,Y) = f(x,y) \quad \text{if } Y = \{y\}, \tag{9}$$

for all sets of bounding boxes $Y \subset \mathcal{Y}$. The condition in Equation (8) guarantees that $\hat{f}$ is a true upper bound, while the condition in Equation (9) ensures global optimality at the time of convergence, by specifying that the bound must be equal to the true function value in the event that the set of bounding boxes contains only one item, y.

In order to apply the ESS optimization to the problem of maximizing kernelized functions of the form given in Equation (6), we first reorder the summation and separate the right hand side into three terms based on their dependence on the restriction kernel, the local context kernel, and the global context kernel, respectively. As we are only interested in the argmax, we can discard the bias term, b, that is a constant independent of x and y.

$$f_{restr}(x,y) = \beta_1 \sum_i \alpha_i k_{I_1}(x|_y, x_i|_{y_i}) \tag{10}$$

$$f_{local}(x,y;\theta) = \beta_2 \sum_i \alpha_i k_{I_2}(x|_{\Theta(y)}, x_i|_{\Theta(y_i)}) \tag{11}$$

$$f_{global}(x,y) = \beta_3 \sum_i \alpha_i k_{I_3}(x, x_i) \tag{12}$$

If we can provide upper bounds for each of these functions, we can upper bound their sum by the sum of their upper bounds.

We note first that f_{global} in Equation (12) has no dependence on y. We can therefore set $\hat{f}_{global}(x,Y) = f_{global}(x,y)$ for arbitrary $y \in Y$, thereby fulfilling both (8) and (9). Next, we

observe that, up to a multiplicative constant, Equation (10) is in exactly the form that was analyzed in [21, 22]. A selection of suitable upper bounds as well as a recipe for constructing bounds with interval arithmetic was given in that work. Therefore, we only need to provide a bound for f_{local}. As for the restriction kernel, this bound will dependent on the image kernel used. In the next section, we discuss the construction of bounds for the concrete example of local context kernels based on a *bag of visual words* representation.

4.1 A Bound for Local Context Kernels

We illustrate the construction of a quality bound for k_{local}, using the visual words kernel discussed in [21]. In this model, we represent image regions by histograms of vector quantized local features, and compute the kernel $k_I(x_i,x_j) = \langle h_{x_i}, h_{x_j} \rangle$, where h_{x_i} is the histogram computed from x_i. Due to the linearity of the kernel, we can rewrite the resulting expression as a sum over individual contributions for each local feature point in a query region:

$$f_{local}(x,y;\theta) = \beta_2 \sum_i \alpha_i k_{I_2}(x|_{\Theta(y)}, x_i|_{\Theta(y_i)}) = \sum_{k \in x|_{\Theta(y)}} w_{c_k} \tag{13}$$

where $w = \beta_2 \sum_i \alpha_i h_{x_i|_{\Theta(y_i)}}$ is a vector of positive and negative per-feature weights, and c_k is the cluster id of the kth local feature point.

Following [21, 22] 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 $Y = \left(\left[\underline{l}, \overline{l}\right], \left[\underline{t}, \overline{t}\right], \left[\underline{r}, \overline{r}\right], \left[\underline{b}, \overline{b}\right] \right)$, we specify intervals

$$\overline{\Theta}(Y) = \left(\left[\underline{l} - \theta\overline{w}, \overline{l} - \theta\underline{w}\right], \left[\underline{t} - \theta\underline{w}, \overline{t} - \theta\underline{h}\right], \left[\underline{r} + \theta\underline{w}, \overline{r} + \theta\overline{w}\right], \left[\underline{b} + \theta\underline{h}, \overline{b} + \theta\underline{h}\right] \right), \tag{14}$$

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. The region of uncertainty for the interior boundary of the local contextual region is simply Y itself. In order to compute an upper bound for which condition (8) holds, we overestimate the number of positive w_{c_k} that will fall in the region $x|_{\Theta(y)}$, for $y \in Y$, and underestimate the number of negative w_{c_k}. We do so by defining four rectangular regions in the image plane: the largest possible rectangle in $\overline{\Theta}(Y)$, which we denote $\overline{\Theta}(Y)_{max}$; the smallest possible rectangle, $\overline{\Theta}(Y)_{min}$; the largest possible rectangle in Y, denoted Y_{max}; and finally the smallest possible rectangle, Y_{min}. For each of these rectangles, we denote the sum of positive weights of features that fall within these regions in the image plane with a superscript $^+$, and the sum of negative weights with a superscript $^-$. With this notation, we can write a valid upper bound compactly as:

$$\hat{f}_{local}(x,Y;\theta) = \overline{\Theta}(Y)^+_{max} + \overline{\Theta}(Y)^-_{min} - Y^+_{min} - Y^-_{max}. \tag{15}$$

Note that this construction automatically fulfills the condition (9): if $Y = \{y\}$, it follows that $\overline{\Theta}(Y)_{min} = \overline{\Theta}(Y)_{max} = \overline{\Theta}(Y)$ and $Y_{min} = Y_{max} = Y$ such that $\hat{f}_{local}(x,Y;\theta) = (\overline{\Theta}(Y)^+ + \overline{\Theta}(Y)^-) - (Y^+ + Y^-) = f_{local}(x,y)$. Using integral images, as in [21], we can compute this upper bound in *constant time*, in particular independently of the number of elements in Y. We now have all the necessary ingredients to apply ESS with global and local context kernels.

4.2 Simultaneous Search Over Multiple Images

For localization within a single image, the global context term does not influence the result returned by ESS. However, if we simultaneously search over multiple images, we can gain from increased computational efficiency. The global context term acts as a prior over images, focusing the search on images that are most likely to contain an object of interest. Unpromising images will have low upper bounds, and will be less likely to arrive at the head of the priority queue. Consequently, fewer splits of the search space will be performed for images with low contextual score, instead focusing the computational effort on more promising images.

5 Experimental Results

To show the performance of the global and local context kernels, we performed experiments on the publicly available PASCAL VOC 2006 and VOC 2007 datasets. In all cases, we represented images by *bag of visual word* histograms and used linear image kernels for k_{restr} and k_{local}, and a χ^2-kernel for k_{global}. Subsequently, we trained a classifier using ground truth bounding box on the training sets as positive examples, and we randomly sampled background regions so that the ratio of positive to negative training data was 1:5. We used the multiple kernel learning algorithm described in [29]. For the detection step, we extended the publicly available version of ESS to incorporate local and global context kernels.[2]

As baseline methods, we compare to the case without context, *i.e.* only the restriction kernel, and the case where the importance of the different context components is fixed instead of learned, by setting $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$. Additionally, we compare to previously reported state of the art results from [21] (Figure 3).

The PASCAL VOC 2006 [11] and VOC 2007 [12] datasets are amongst the most difficult datasets currently in use to benchmark object classification and object localization systems. For our experiments, we extract local SURF descriptors [5] from interest points, random positions and on the regular grid, and quantize them into a 3000-bin *bag of visual word* histogram using a codebook that was created using k-means clustering.

Because we are interested in *localization performance* only, we follow the evaluation procedure proposed in [21] instead of one used in the VOC challenges, as those yield a combined measure of classification and localization performance. To evaluate the performance for any of the 10 object categories (2006) or 20 object categories (2007), we measure the *average precision (AP)* score (see [11]) achieved on only the test images that actually contain the current object class. Table 1 summarizes the resulting scores in numeric form. Additionally, Figure 3 shows the precision recall curves for the VOC 2006 categories *cat* and *dog* that were also reported in [21].

From the tables one can see that the use of context consistently improves the localization quality (here and in the following we disregard the 8 categories where all methods have AP scores below 0.1, because the measure is dominated by random effects in this range). For 19 of the 22 relevant categories, the global and local context kernels with learned weights achieved better results than the equally weighted version. Compared to the setup without context, one achieves an improvement in average AP scores from 0.17 to 0.29 when performing localization on VOC 2006 with learned context weights (0.22 for fixed averaging), and from 0.14 to 0.23 on VOC 2007 (0.17 for fixed averaging). Figure 3 supports this

[2] Source code will be made available at the time of publication.

	bicycle	*bus*	*car*	*cat*	*dog*	*cow*	*horse*	*m.bike*	*person*	*sheep*
learned	0.410	**0.253**	**0.268**	**0.415**	**0.332**	**0.286**	**0.206**	**0.413**	0.049	**0.229**
fixed	**0.429**	0.177	0.263	0.251	0.178	0.194	0.167	0.344	0.015	0.182
no context	0.396	0.100	0.145	0.259	0.170	0.118	0.165	0.276	0.036	0.027

	aeroplane	*bicycle*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*
learned	**0.114**	**0.122**	0.088	0.060	0.000	**0.254**	0.140	**0.356**	0.091	**0.137**
fixed	0.105	0.115	**0.123**	0.049	0.003	0.252	**0.150**	0.231	0.091	0.119
no context	0.050	0.064	0.069	0.036	0.026	0.106	0.068	0.312	0.019	0.121

	d.table	*dog*	*horse*	*m.bike*	*person*	*p.plant*	*sheep*	*sofa*	*train*	*tv*
learned	**0.242**	**0.321**	**0.273**	**0.344**	0.026	0.016	0.091	**0.301**	**0.295**	0.078
fixed	0.030	0.220	0.262	0.187	0.028	0.013	0.091	0.131	0.292	0.045
no context	0.147	0.228	0.153	0.165	0.091	0.022	0.091	0.160	0.193	0.030

Table 1: Average precision on PASCAL VOC 2006 (top) and VOC 2007 (bottom) dataset using local and global context kernel with *learned* weighted, *fixed* weights and with *no context* (only restriction kernel). We indicate the best result for each category by bold print, except for AP scores below 0.1, which we do not consider significant, as the AP measure is very unstable in this regime.

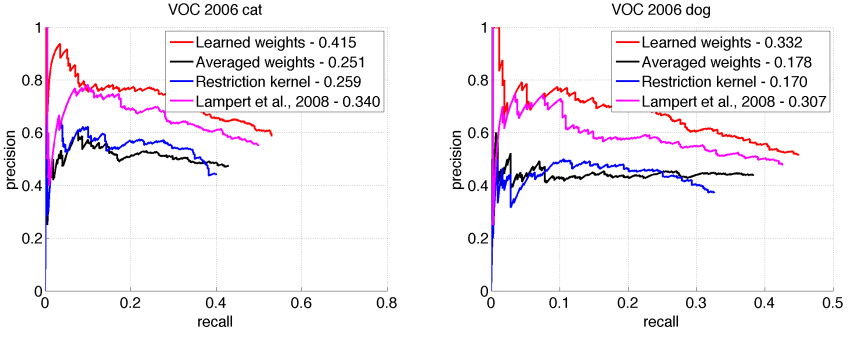


Figure 3: Precision-Recall curves for categories *cat* and *dog* of PASCAL VOC 2006.

observation: localization with the (context-unaware) restriction kernel and with the fixed averaging context-aware kernel achieve approximately the same performance in this case. The localization with learned weights achieves clearly higher precision and recall than both. Its results also improve over the best results reported in the literature so far (Lampert et al. [21]). Note that this work uses the same restriction kernel that we use as baseline, but with a different ranking function.

Note that it is really the per-class selection of context weights that has this positive effect. Averaged over all classes, the coefficients $\beta = (\beta_{restr}, \beta_{local}, \beta_{global})$ are $(0.48 \pm 0.13, 0.39 \pm 0.10, 0.24 \pm 0.21)$ for VOC 2006 and $(0.30 \pm 0.16, 0.31 \pm 0.19, 0.39 \pm 0.32)$ for VOC 2007. This shows that, overall, all kernels are of roughly equal importance. Therefore, it is the significant variations that occur for the different classes that cause the positive effect on localization accuracy.

6 Conclusions

In this work, we have proposed a method for the integration of local and global context into kernel-based classifiers (SVMs) that can learn the importance of different context contribu-

tions efficiently as part of the training procedure. The *local and global kernels* framework combines the advantages of efficient inference with rectangular context regions with the ability to learn the importance of different context contributions from the training data. Experiments on the PASCAL VOC 2006 and VOC 2007 datasets showed that the ability to adapt the context influence to the target class at hand is a crucial factor in order to benefit from the use of context.

The flexibility of *local and global kernels* makes possible several future extensions. As it is generally accepted that the use of more training data and the combinations of more feature types improves the performance of object localization systems, we presume that the use of more than two context kernels in Equation 4 will improve the performance as well. Additionally, the spatial extent of informative contextual regions is likely to be class dependent, suggesting that this should be included in the learning procedure. An interesting aspect would also be the extension to other shapes than just rectangular context, *e.g.* the use of superpixel segmentations. Finally, it has recently been shown that *structured regression training* can lead to improved detection even for otherwise identical setups [2, 7, 19]. We plan to make use of this in future work.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007- 2013) / ERC grant agreement no. 228180. This work was funded in part by the EC project CLASS, IST 027978, and the PASCAL2 network of excellence. The first author is supported by the Royal Academy of Engineering through a Newton International Fellowship.

References

[1] E. H. Adelson. On seeing stuff: The perception of materials by humans and machines. In *Proceedings of the SPIE*, volume 4299, 2001.

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

[3] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 2004.

[4] A. Baumgartner, C. Steger, H. Mayer, and W. Eckstein. Multi-resolution, semantic objects, and context for road extraction. In *SMATI*, 1997.

[5] Herbert Bay, Tinne Tuytelaars, and Luc J. Van Gool. SURF: Speeded up robust features. In *ECCV*, pages 404–417, 2006.

[6] N. H Bergboer, E. O. Postma, and H. J. Herik. Context-based object detection in still images. *Image and Vision Computing*, 24(9), 2006.

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

[8] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.

[9] N. Dalai, B. Triggs, I. Rhone-Alps, and F. Montbonnot. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[10] S.K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009.

[11] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[13] M. Fink and P. Perona. Mutual boosting for contextual influence. In *NIPS*, 2003.

[14] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3), 2008.

[15] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.

[16] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[17] H. Kruppa and B. Schiele. Using local context to improve face detection. In *BMVC*, 2003.

[18] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005.

[19] C. H. Lampert and M. B. Blaschko. Structured prediction by joint kernel support estimation. *Machine Learning*, 2009.

[20] C. H. Lampert and M. B. Blaschko. A multiple kernel learning approach to joint multi-class object detection. In *DAGM*, volume 5096. Springer, 2008.

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

[22] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *Pattern Analysis and Machine Intelligence*, 2009.

[23] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5, 2004.

[24] S. Lazebnik and M. Raginsky. An empirical bayes approach to contextual region classification. In *CVPR*, 2009.

[25] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman. *Remote sensing and image interpretation*. Wiley, 2004.

[26] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. In *NIPS*, 2003.

[27] S. E. Palmer. The effect of contextual scenes on the identification of objects. *Memory and Cognition*, 3, 1975.

[28] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.

[29] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9, 2008.

[30] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR*, 2007.

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

[32] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

[33] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. In *CVPR*, 2003.

[34] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2), 2003.

[35] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *ICCV*, 2003.

[36] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, 2004.

[37] M. M. Tuffield, S. Harris, D. P. Dupplaw, A. Chakravarthy, C. Brewster, N. Gibbins, K. O'Hara, F. Ciravegna, D. Sleeman, N. R. Shadbolt, and Y. Wilks. Image annotation with photocopain. In *Semantic Web Annotation of Multimedia (SWAMM)*, 2006.

[38] J. R. R. Uijlings, A. W. M. Smeulders, and R. J. H. Scha. What is the spatial extent of an object? In *CVPR*, 2009.

[39] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 69(2), 2006.