Combining efficient object localization and image classification

Hedi Harzallah, Frédéric Jurie, Cordelia Schmid

Combining efficient object localization and image classification

Hedi Harzallah Frédéric Jurie Cordelia Schmid
LEAR, INRIA Grenoble, LJK
firstname.lastname@inrialpes.fr

Abstract

In this paper we present a combined approach for object localization and classification. Our contribution is twofold. (a) A contextual combination of localization and classification which shows that classification can improve detection and vice versa. (b) An efficient two stage sliding window object localization method that combines the efficiency of a linear classifier with the robustness of a sophisticated non-linear one. Experimental results evaluate the parameters of our two stage sliding window approach and show that our combined object localization and classification methods outperform the state-of-the-art on the PASCAL VOC 2007 and 2008 datasets.

1. Introduction

Over the past years, there has been increasing interest in object category recognition. Two major tasks are image classification and object localization. *Image classification* is defined as the task of assigning an image one or multiple labels corresponding to the presence of a category in the image. Many recent papers exist on this topic [19, 23, 27, 30] and several evaluation campaigns, such as TRECVID, Pascal VOC, and ImageCLEF, demonstrate the good performance of these algorithms. *Object localization* detects instances of a given category in the image, in many cases up to a bounding box. Recent techniques combine efficient image descriptors with machine learning techniques [5, 7, 12, 13, 18, 28]. The task remains challenging due to intra-class variations, viewpoint changes and deformations of the objects.

Even if the two tasks are different, they are obviously related. If one has a good object detector, it becomes easy to predict image labels when some objects are detected with high scores (Figure 1-a) . Inversely, knowing the class of an image can help to detect hardly visible objects (Figure 1-b). This interdependence has been studied in context-based approaches, where context can model properties of the entire image [14, 26], the relationship between objects [4, 16] or the background surrounding the candidate detections [6]. Convincing results have been obtained by approaches where the image is segmented into labeled regions and the labels are used for the localization of surrounding objects [13, 29]. A recent paper [8] presents an empirical evaluation of the role of context in object detection, evaluating several sources of different context and ways to utilize it.

(a) (b)

Figure 1. Complementarity of image classification and object localization: (a) Cars are detected with strong scores, while the image obtained a poor classification score. (b) Cars are detected with low scores due to partial visibility, but the image got a very high classification score.

Because of the interdependence of the two tasks, the combination of object detection and image classification became a topic of interest in the recent literature. Li and Fei-Fei [20] propose a graphical model of events in images where *event* is a latent factor conditioning the generation of objects and scene categories. Shotton *et al.* [25] propose a related idea in the context of image segmentation: the likely categories are emphasized by multiplying the local segmentation and global image classification distributions. This is a principle we find again in [13], in another context, for improving object detections using labels of surrounding regions. We build on the idea that classification and detection can be considered as independent knowing a latent property of the observed scene, leading to a simple but powerful combination scheme. The main contribution of this paper is to show that such a combination can improve the results of state-of-the-art algorithms, both for classification and detection.

This, obviously, requires state-of-the-art approaches for classification and localization. For image classification we

rely on [22], one of the state-of-the-art approaches of the PASCAL VOC 2007 and 2008 challenges. For object localization we build on and improve existing sliding window approaches [1, 5, 12, 24]. Firstly, we implement an efficient two stage approach which uses a linear support vector machine (SVM) classifier for pre-selection and a non-linear SVM for scoring. This allows an excellent trade-off between speed and accuracy. We believe that our two stage cascade is simpler than recent cascades [2, 3, 9] while giving remarkably good results. Secondly, we evaluate extensively different ways of describing the image and propose a simple and efficient image representation. Our window descriptor builds on recent work [1, 5, 7, 19] and combines the key ideas of these approaches into a simple but efficient descriptor. A comparison with the state-of-the-art shows that our detector gives better results for most of the object categories. This detector represents the second contribution of the paper.

The paper is organized as follow. Section 2 presents the datasets used for our experiments. A description of our efficient object detector and its experimental evaluation is given in section 3. Section 4 then describes the model we propose for combining classification and localization and evaluates its performance. Finally, in section 5 we present a comparison with the state-of-the-art methods and conclude.

2. Dataset and evaluation criteria

The PASCAL visual object class datasets are today probably the most widely used reference datasets for category recognition. Objects are present in realistic conditions under scale, viewpoint and illumination changes as well as with significant amounts of background clutter, see figure 4 for an illustration. The PASCAL VOC 2007 & 2008 datasets [10] used in this paper contain twenty object classes: person, animals (bird, cat, cow, dog, horse, sheep), vehicles (aeroplane, bicycle, boat, but, car, motorbike, train), and indoor objects (bottle, chair, dining table, potted plant, sofa and TV/monitor).

The main PASCAL challenge tasks are image classification and object localization. Training is in both cases supervised. For object localization the annotations include the bounding box, the object pose (left, right, front, back, other) and a flag indicating whether the object is truncated (only part of the object is visible). Training, validation and test sets are available for PASCAL VOC 2007, but not for 2008 for which the test annotations have not yet been made available.[1] We, therefore, run our evaluations in sections 3 and 4 on the 2007 dataset. We, then, compare in section 5 to the state-of-the-art on the VOC 2007 and 2008 datasets.

Our performance metrics, both for classification and detection, follow the PASCAL VOC ones. The *average precision (AP)* is computed from the precision/recall curve and is an approximation of the area under this curve. The mean AP (mAP) measures the mean of the APs over all categories.

[1] An evaluation of the results on the VOC 2008 test set can be obtained from M. Everingham.

3. Efficient object localization

Our object localization approach builds on the now standard sliding window approach [7, 12, 18, 28]. Such an approach evaluates a score function for all positions and scales in an image and detects local maxima of this function. Its performance depends on: (a) an efficient *search strategy*; (b) a robust *image representation*; (c) an appropriate *score function* for comparing candidate regions with object models; (d) a *multi-view representation* and (e) a reliable *non-maxima suppression*.

Our approach has two contributions: a robust window representation and an efficient search strategy. For all the remaining components we use the standard techniques. Our two contributions are described and evaluated below.

3.1. Image representation

Our approach uses two complementary descriptors, shape and appearance descriptors described in the following. The combination of the two descriptors is presented and evaluated in the next section.

Shape descriptor (HOG). The shape descriptor is a histogram of oriented gradients (HOGs) [7]. We apply the fast HOG implementation of [31]: after quantizing the gradient orientation at each pixel, we compute and store an integral image for each discrete orientation. These integral images are used to efficiently compute the HOG for any rectangular image region, i.e., in $4 \times \texttt{number of orientations}$ basic operations.

The division of the description window into sub-windows (or *tiles*) is referred to as geometry. The geometry is determined by three main parameters: (1) The number of tiles that determine the resolution of the tiling. (2) The organization of the tiles which can be *adapted* or *regular*. Adapted tiles are as square as possible and are obtained as follows: given the average aspect ratio of the object category and the number of desired tiles T, we seek to have tiles that are as square as possible with $round(\sqrt{\frac{TW}{H}})$ tiles along the width, and $round(\sqrt{\frac{TH}{W}})$ along the height. Regular tiles are taken on a grid with the same number of tiles along the height and the width. (3) The overlap between tiles: when overlapping, a tile shares 50% of its surface with each of its four neighbors .

We build different configurations with the number of tiles ranging from 40 to 350, with overlapping or not overlapping tiles, and with adapted or regular grids. We evaluate these configurations with the filtering classifier (linear

kernel). To compare the results obtained with the different configurations, we perform the *Friedman* statistical test. This test, based on the ranks on the different classes, separates the configurations into significantly different sets (here with probability of 95%).

We obtained 12 groups of equivalent configurations and observed that the difference between the best and the worst one results in a significant difference in mAP (7%). Three main observations are: (1) Tilings should contain at least 150 tiles, but having more has no impact. (2) Tiles should be overlapping; we observed a difference up to 3% just by adding the overlap between tiles. (3) Tiles should be as square as possible, but the difference is less significant.

We also evaluated the influence of the number of discrete HOG orientations. We used several configurations where the number of bins varied from 8 to 32 with signed or unsigned orientations. Using signed orientations appeared to be better ($+1.3\%$), 16 bins outperformed 8 bins by 3% and 32 bins lead to the same performance as 16.

Appearance descriptor (BOF). The appearance descriptor builds on the spatial pyramid over quantized SIFT descriptors [19]. We first extract multi-scale patches from the image and describe them using the SIFT descriptor [21]. The descriptors are then quantized into visual words using k-means and a histogram of visual words summarizes the content of the window. Instead of building one global histogram for the window, we compute one histogram per tile, using the spatial pyramid tiling. This technique consists in partitioning the image into increasingly finer subregions and computing histograms of local features inside each subregion.

We vary the level of the pyramid and calculate the mAP on the PASCAL VOC 2007 dataset. Using only one level leads to poor results, the mAP is less than 3%. Using two levels rises the mAP to 7.6% and the best configuration uses 3 levels, resulting in a mAP of 15%. Adding more levels does not improve the results further. These experiments were obtained with a linear classifier and a visual vocabulary of 100 words.

We observe that the BOF descriptor performs slightly better than the HOG (14.6% mAP against 15%, see table 1). On the other hand, the BOF descriptor with 3 levels is more expensive computationally as well as memory-wise.

Normalization of the descriptors. Normalization of the descriptors makes them robust to contrast changes (HOG) or to scale changes (BOF and HOG). Previous work [7, 31] showed that local L1 and L2 norms give comparable results. In the following, we use the L2 norm, which can be used in different ways. We can either normalize each cell or perform a global normalization of the descriptor. It is also possible to not normalize the descriptor at all. Experimental results show that (a) not using any normalization reduces the results by 14% (HOG) and 4% (BOF) and (b) the per tile normalization is about 2% better than global normalization for both HOG and BOF.

3.2. Two stage object localization

3.2.1 Search strategy

As the number of windows per image is huge, a technique for reducing this number should be applied. While some authors suggest methods avoiding to scan exhaustively the image [5, 17], the most popular technique is the *cascade*, introduced by Viola and Jones [28]. It decomposes a strong classifier into several classifiers arranged in a cascade, each of which decides if the window contains the object or not. This type of approach has received a lot of attention during the past five years [2, 3, 9, 31]. Cascades have, however, several limitations. Training a cascade is slow, taking on the order of weeks; determining the target false positive rate and detection rate at each stage in the cascade is often empirical; and finally, a cascade always reduces the overall performance.

We propose to use a very simple but efficient two stage cascade. We first apply a linear SVM for each window of the image. This is fast due to the simplicity of the classifier and the use of fast descriptors implemented with integral images. We then apply the final score function (a non-linear SVM with a χ^2 kernel) only on candidate regions, i.e., those that obtained good scores during the first stage. Our experimental results confirm that this approach has an excellent trade-off between speed and accuracy.

3.2.2 Linear classifier

The filtering classifier is a linear SVM classifier that is used to rapidly scan the image and reject windows unlikely to contain objects. Its performance depends on the data used to train it. We artificially increase the positive training set by following a procedure proposed by Laptev [18]. The negatives examples are obtained by an iterative procedure. The initial training set consists of randomly chosen background windows and objects from other classes. The resulting classifier is used to scan images and select the top false positives or *hard examples*. These hard examples are added to the negative set and a new classifier is learned. This procedure is repeated several times to obtain the final classifier.

The performance of the filtering classifier is reported in Table 1. The mAP is 14.6% for HOG features and increases by 3% when combining HOG and BOF features. The combination is obtained by concatenating the two feature vectors.

However, the main interest of a filtering classifier is its ability to filter out the majority of the windows, the scoring classifier being applied only on the remaining small fraction of these windows. It is, therefore, more important to measure the capability of the linear classifier to select windows

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP
Linear, HOG	10.0	27.8	04.7	00.6	11.4	31.7	33.9	02.6	10.1	14.9	09.7	01.8	28.1	22.6	12.2	09.9	10.0	04.3	19.3	26.1	14.6
Linear, BOF	16.9	21.2	04.9	04.8	07.3	25.2	28.4	06.9	09.8	10.3	06.7	06.9	30.5	26.6	13.1	09.4	12.5	12.1	17.0	28.5	15.0
Linear, HOG+BOF	22.4	30.5	03.3	01.8	11.2	26.4	36.7	06.0	11.1	14.3	10.9	07.6	33.8	27.2	14.7	09.8	15.1	14.7	22.4	32.2	17.6
χ^2, HOG	18.4	39.5	09.8	02.0	18.2	42.2	47.5	02.5	13.6	22.1	10.5	10.7	43.5	34.6	14.5	11.7	12.7	14.2	31.8	37.2	21.9
χ^2, BOF	29.8	33.3	11.1	04.2	09.5	39.7	42.3	14.4	12.7	20.4	13.3	15.5	40.5	37.6	16.8	11.4	19.8	18.8	34.4	35.6	23.1
χ^2, HOG+BOF	33.8	43.0	09.7	09.6	18.7	41.9	50.4	15.0	14.6	23.9	15.1	15.4	48.2	41.7	20.2	16.1	21.2	20.3	29.1	38.2	26.3

Table 1. Localization performance on PASCAL VOC 2007 for the different image representations with/without using the non-linear scoring classifier.

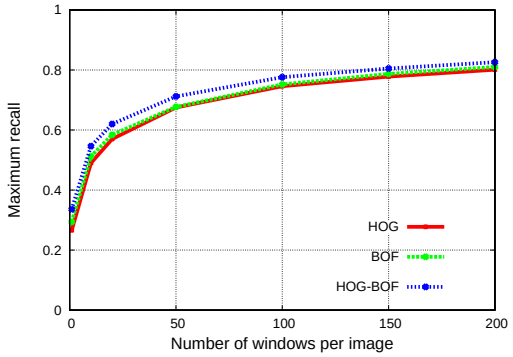

Figure 2. Maximum recall versus number of windows per image for filtering with a linear SVM with HOG and/or BOF features. Results are presented for PASCAL VOC 2007.

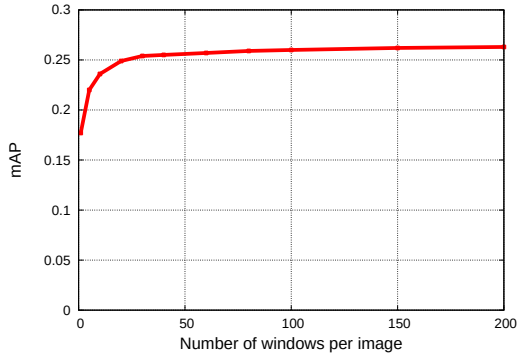

Figure 3. mAP of the scoring classifier (χ^2 kernel + SVM with HOG & BOF features) versus number of windows per image selected by the filtering classifier (linear SVM and HOG features). Results are presented for PASCAL VOC 2007.

than to measure its mAP, even if the two are related. We measured this capability with two criteria: (1) The highest recall the scoring classifier can reach using the top N windows selected by the linear classifier, for different values of N. Results are given Figure 2. (2) The mAP obtained by the scoring classifier if only the top N windows are selected. These results are reported in Figure 3 for filtering with a linear SVM + HOG and scoring with χ^2 SVM with HOG & BOF features.

From these experiments, we can draw two conclusions. First, we observe that the different possible descriptors (HOG, BOF or the combination of the two) give similar results in terms of the highest possible recall, see Figure 2. Therefore, we use in the filtering stage the least expensive descriptor, i.e., HOG. Second, the mAP does not progress significantly for value above $N = 100$.

3.2.3 Improvement due to non-linear classifier

The scoring classifier is based on a non-linear SVM with a χ^2 kernel: $K(x, v) = exp(-\gamma_{ch}.d_{\chi^2}(x_{ch}, v_{ch}))$ where γ_{ch} is a channel dependant parameter, x_{ch} and v_{ch} are the components of vectors x and v corresponding to the channel ch, here shape (HOG) or appearance (BOF), and d_{χ^2} is the χ^2 distance defined by

$$d_{\chi^2}(x, v) = \sum_{i=1}^{N} \frac{(x_i - v_i)^2}{|x_i + v_i|}. \quad (1)$$

The γ value is defined by the heuristic of [30] as follows:

$$\gamma = \frac{Nb^2}{\sum_{i=1}^{Nb} \sum_{j=1}^{Nb} d_{\chi^2}(x_i, x_j)}, \quad (2)$$

where Nb the number of training examples and x_i the i^{th} training example.

To train the scoring classifier, we use the same positive training examples as for the filtering classifier. For the negative examples, we do not find additional hard examples due to the computational cost, but use the negative examples of the filtering classifier (background windows and hard examples) to which we add a large number (here 70K) of background windows.

Table 1 shows the performance increase due to the non-linear scoring stage. The table reports the results obtained using a χ^2 kernel and three different descriptors combinations when filtering with a linear SVM + HOG and keeping 200 windows per image. The best mAP (26.3%) is obtained for the combination of HOG and BOF. Using only one of

these representations reduces the performance by more than 3%. The improvement due to the non-linear classifier is very significant (8.7%).

Note that there is also an increase of performance when combining HOG and BOF for a linear SVM classifier. The combination of the two features allows to improve detection score, but does not help to improve the filtering efficiency of the linear SVM, see figures 2 and 3.

3.3. Discussion

Our final detector, i.e., χ^2, HOG + BOF, gives excellent results on the PASCAL VOC 2007 database (see Table 1, bottom line). The proposed detector is efficient due to the two stage sliding window algorithm and its results compare favorable to the state-of-the-art. Experimental results show the benefit of using complementary image descriptors. Figure 4 shows a few localization examples. The first row shows two correctly detected objects, while the second row shows two strong false positives. Interestingly, the detector confuses a bus with a car and includes context information in the case of chairs. The last row shows missed objects and illustrates the complexity of the task: the car is very small and the chair hardly visible.

4. Contextual combination of localization and classification

In the following we describe our approach for combining localization and classification and present an experimental validation.

4.1. Our approach

The idea that classification and detection can benefit from and contribute to each other's successes relies on the assumption that they use different information. This assumption can be verified by observing that, often, a single image is classified differently (i.e. with a significantly different probability) by the detector and the classifier. Figure 5 shows the probability density of true positives windows (all classes of the Pascal VOC 2007 dataset being merged) as a function of the probabilities given by the classifier (applied to the whole image) and the detector (applied to the window). Interestingly, we can see that our hypothesis is valid: for many true positives only one of the modalities has a high probability. Indeed, if an object is occluded or truncated, it will be hardly detectable by the detector while the classifier could still have enough information (context, object parts) to decide on the presence of the object. Inversely, if the object is small and appears in a non standard context, the detector will still be able to find it while it would be non-detectable for the classifier (see Figure 1 for an illustration). We also observed (not shown on the figure) that for most false positives both probabilities are low.

In the following, we construct a model for combining lo-

Figure 4. Localization examples for the classes car and chair with our detector (χ^2, HOG + BOF) on PASCAL VOC 2007. (a) Examples of objects correctly detected. (b) Examples of false positives with high score. (c) Examples of objects that were missed.

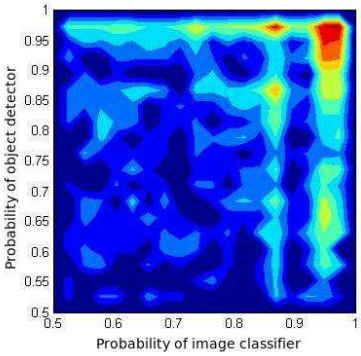

Figure 5. Probability density of true positive windows, as a function of the probabilities given by the image classifier and the object detector. Warm (resp. cold) colors mean high (resp. low) densities. Results are presented for the PASCAL VOC 2007 dataset.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP
Localization	58.2	59.5	19.3	26.9	37.5	60.8	80.2	36.7	39.3	40.7	28.3	35.2	67.9	63.8	74.0	38.4	40.1	35.7	56.7	54.3	47.7
INRIA_Flat_V2	76.7	64.4	55.8	**69.3**	34.1	62.6	76.0	**58.4**	55.9	44.3	60.3	**48.3**	77.4	63.5	85.7	42.9	48.3	49.0	**76.4**	54.5	60.1
Product	75.0	69.0	51.7	67.7	**48.8**	**68.7**	83.3	54.6	57.6	**53.7**	56.6	46.4	**78.7**	69.4	84.8	51.8	**54.4**	**55.5**	73.0	**62.8**	63.2
Our combination	**77.2**	**69.3**	**56.2**	66.6	45.5	68.1	**83.4**	53.6	**58.3**	51.1	**62.2**	45.2	78.4	**69.7**	**86.1**	**52.4**	**54.4**	54.3	75.8	62.1	**63.5**

Table 2. Image classification performance on the PASCAL VOC 2007 dataset obtained by localization and classification methods (1st and 2nd rows) and by combination (last 2 rows).

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP
Our detector	33.8	43.0	09.7	09.6	18.7	41.9	50.4	15.0	14.6	23.9	15.1	15.4	48.2	41.7	20.2	16.1	21.2	20.3	29.1	38.2	26.3
Context of [11]	**35.4**	44.7	10.6	05.5	20.3	**42.1**	50.7	18.7	16.2	26.2	14.2	16.3	49.0	42.9	20.5	17.0	23.4	21.0	**30.7**	39.3	27.2
Product	33.9	40.7	10.8	11.9	22.7	36.8	47.1	18.7	17.4	**31.9**	17.0	16.5	48.0	38.1	20.9	18.4	26.6	23.6	28.6	37.2	27.3
Our combination	35.1	**45.6**	**10.9**	**12.0**	**23.2**	**42.1**	**50.9**	**19.0**	**18.0**	31.5	**17.2**	**17.6**	**49.6**	**43.1**	**21.0**	**18.9**	**27.3**	**24.7**	29.9	**39.7**	**28.9**

Table 3. Object localization performance on the PASCAL VOC 2007 dataset obtained by our two stage detector alone (1st row) and combined with context (last 3 rows).

calization and classification results that takes into account this notion of *detectability*. Let us denote $P(D_i)$ as the probability that the presence of an object can be detected by the classifier applied to the entire image. In the same way, $P(D_w)$ is the probability that the presence of an object can be detected by the sliding window detector. We obtain an approximation of the conditional probability $P(O|D_i, S_i)$, where S_i is the score of the classifier, by histogramming the scores of positive and negative training examples. Objects of all training images are assumed to be detectable. This gives us the probability of having the object in the image knowing it is detectable. In the same way, we compute an approximation of $P(O|D_w, S_w)$, the probability of having an object knowing the score S_w of the sliding window detector and assuming it is detectable.

Formally, the probability of having an object in the image, given the classification score, is therefore

$$P(O|S_i) = P(D_i)P(O|S_i, D_i) + P(\overline{D_i})P(O|S_i, \overline{D_i}) \quad (3)$$

and similarly, the probability of having an object in a window given the detection score is

$$P(O|S_w) = P(D_w)P(O|S_w, D_w) + P(\overline{D_w})P(O|S_w, \overline{D_w}) \quad (4)$$

where $P(O|S_i, \overline{D_i})$ (resp. $P(O|S_i, \overline{D_w})$) is the probability that the object is present when it is not detectable by our image classifier (resp. object detector). They are supposed here to be constant values.

The final probability is task dependent. For the localization task, we consider the score S_w of the window obtained by the detector as well as the score S_i of the image obtained by the classifier, and compute the probability of having an object in the window as:

$$P(O|S_w, S_i) \propto P(O|S_i) \times P(O|S_w). \quad (5)$$

For the classification task, we consider the score of the window having the best score, denoted S_{bw}, as well as the score S_i obtained by the image classifier. The probability of having an object in the image is then computed as:

$$P(O|S_{bw}, S_i) \propto P(O|S_i) \times P(O|S_{bw}). \quad (6)$$

For our experiments, conditional probabilities $P(O|S_i, \overline{D_i})$, $P(O|S_i, \overline{D_w})$ as well as the priors $P(D_i)$ and $P(D_w)$ are constant values obtained by cross-validation, i.e., we take the values that maximize the AP on a validation set.

4.2. Experiments

Classification experiments. We compare the classification performance obtained by our detection framework, by the INRIA_Flat_V2 approach and by the combination of both. To obtain classification scores based on detection we keep for each object class only the best scored window in the image. The INRIA_Flat_V2 approach [22], for which we have obtained the results (classification scores) from the authors, is based on the work of [30], i.e., it integrates a set of different image features with a χ^2 kernel. V2 signifies the version used in the PASCAL VOC 2008 challenge which adds additional channels to the 2007 version and improves the performance. Table 2 shows the results for the PASCAL VOC 2007 dataset. The first row is obtained by detection. In terms of mean average precision (mAP), it gives poor results. However, for some classes it gives very good results and even outperforms the INRIA_Flat_V2 (second row) approach, namely for the classes car and bottle.

When combined ("Our combination"), we observe a significant gain of 3.4% in term of mAP over the INRIA_Flat_V2 classifier. Furthermore, for some classes (e.g bottle, plant) we observe a gain of 10% in AP which represents a significant improvement. The row "Product" shows

the results obtained with a simpler combination rule where we multiply the two probabilities. As expected, the results are not as good as with our combination method. We also evaluated other combination methods such as the MIN or MAX rules [15], and an SVM classifier for a prediction based on two scores. All of them performed slightly worse than our approach.

Detection experiments. We compare the results obtained by our detector to combinations with the classification score given by the INRIA_Flat_V2 classifier. Table 3 presents the results. The first row gives the average precision obtained by our detector, the third one shows results for a combination performed by simply computing the product of the two probabilities, and the last one present the results of our method. Our method improves the results of the base detector by 2.6%. Best improvement are observed for the animal classes cow (7.6%) and sheep (6.1%) as well as for indoor classes bottle (4.5%) and chair (3.4%) which suggests that the context information offers most of the improvement.

The results submitted by the authors of [11] to the PASCAL VOC 2008 used contextual information to reweight detection scores: the final score is computed from the initial score, the position of the window and the best detection scores in the image for all the other object categories. The results given by this approach, reported Table 3-row 2, show that the information due to the image classifier is more useful than the contextual information due to object detections of other categories.

5. Comp. with state-of-the-art & Discussion

As stated in the introduction, we believe that to clearly demonstrate the importance of contextual information or of any combination of methods we need to experimentally demonstrated that it produces results better than existing state-of-the-art methods. Starting from any baseline algorithm and improving its performance is indeed easy. Surpassing state-of-the art results is more difficult.

Classification. Table 4 compares our method to the five top methods of the PASCAL VOC 2008 challenge [10]. The mAP we obtain is 57.7%, which improves by 2.8% over the best competing approach. Furthermore, our approach obtains best results for 13 out of 20 categories. On the PASCAL VOC 2007 dataset we obtain 63.5%, see table 2. Compared to the results obtain in the 2007 challenge, we improve by 4.1% in terms of mAP over the best method and obtain best performance on 15 out of 20 classes.

Localization. We compare our localization performance to the best results of the PASCAL VOC 2008 challenge [10] in Table 5. Compared to these results, we achieve best performance for 11 out of 20 classes and obtain results comparable with the method [11].

On the PASCAL 2007 VOC dataset, we obtain 28.9% mAP, see table 2. When compared to the challenge results in 2007, we obtain the best performance for 18 out of 20 classes and a gain of 11.8% in mAP with respect to the best method. This confirms the big improvement of localization methods over the past year.

Discussion. This paper has shown that there exists a potential for combining the presence and location of objects, resulting in a significant improvement of classification and localization performance. We also introduced an efficient two stage sliding window detector. Future work will explore the combination of different types of information. We will, for example, investigate image classification methods that explicitly take into account the results of an object detector, i.e., for example by griding the image based on the detected region.

Acknowledgements This research has been funded by MBDA.

References

[1] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.

[2] L. Bourdev and J. Brandt. Robust object detection via soft cascade. In *CVPR*, 2005.

[3] S. Brubaker, J. Wu, J. Sun, M. Mullin, and J. Rehg. On the design of cascades of boosted ensembles for face detection. *IJCV*, 77(1-3):65–86, 2008.

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

[5] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007.

[6] D. Crandall and D. Huttenlocher. Composite models of objects and scenes for category recognition. In *CVPR*, 2007.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] 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.

[9] M. Dundar and J. Bi. Joint optimization of cascaded classifiers for computer aided detection. In *CVPR*, 2007.

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge Results. http://www.pascal-network.org/challenges/VOC/.

[11] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP
LEAR_flat	80.1	51.8	60.5	66.9	29.1	52.0	57.4	58.6	48.7	31.0	39.2	47.6	64.2	64.6	87.0	28.6	33.3	42.6	73.1	59.8	53.8
LEAR_shotgun	**81.1**	52.9	**61.6**	67.8	29.4	52.1	58.7	59.9	48.5	32.0	38.6	47.9	65.4	65.2	87.0	29.0	34.4	43.1	74.3	61.5	54.5
SurreyUvA_SRKDA	79.5	54.3	61.4	64.8	30.0	52.1	59.5	59.4	48.9	**33.6**	37.8	46.0	66.1	64.0	86.8	29.2	**42.3**	44.0	**77.8**	61.2	54.9
UvA_Soft5ColorSift	79.7	52.1	61.5	65.5	29.1	46.5	58.3	57.4	48.2	27.9	38.3	46.6	66.0	60.6	87.0	31.8	42.2	45.3	72.3	**64.7**	54.0
UvA_TreeSFS	80.8	53.2	**61.6**	65.6	29.4	49.9	58.5	59.4	48.0	30.1	39.6	45.0	**67.3**	60.4	87.1	30.1	41.5	45.4	74.3	59.8	54.3
Our method	80.1	**57.7**	60.7	**69.4**	**43.7**	**52.7**	**70.0**	**60.7**	**50.9**	33.4	**42.9**	**50.1**	66.0	**69.3**	**87.3**	**36.1**	40.4	**45.5**	73.5	64.4	**57.7**

Table 4. Classification performance compared to the state-of-the-art on the PASCAL VOC 2008 dataset.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP
CASIA_Det	25.2	14.6	9.8	10.5	6.3	23.2	17.6	9.0	9.6	10.0	13.0	5.5	14.0	24.1	11.2	3.0	2.8	3.0	28.2	14.6	12.7
Jena	4.8	1.4	0.3	0.2	0.1	1.0	1.3	-	0.1	4.7	0.4	1.9	0.3	3.1	2.0	0.3	0.4	2.2	6.4	13.7	-
MPI_struct	25.9	8.0	10.1	5.6	0.1	11.3	10.6	**21.3**	0.3	4.5	10.1	**14.9**	16.6	20.0	2.5	0.2	9.3	12.3	23.6	1.5	10.4
Oxford	33.3	24.6	-	-	-	-	29.1	-	-	12.5	-	-	32.5	34.9	-	-	-	-	-	-	-
UoCTTIUCI	32.6	**42.0**	**11.3**	11.0	**28.2**	23.2	32.0	17.9	**14.6**	11.1	6.6	10.2	32.7	38.6	**42.0**	**12.6**	16.1	13.6	24.4	**37.1**	**22.8**
XRCE_Det	26.4	10.5	1.4	4.5	0.0	10.8	4.0	7.6	2.0	1.8	4.5	10.5	11.8	13.6	9.0	1.5	6.1	1.8	7.3	6.8	07.1
Our method	**36.6**	33.8	10.7	**11.4**	23.3	**23.7**	**36.6**	15.8	12.9	**17.9**	**15.1**	9.7	**36.5**	**39.4**	19.6	11.6	**19.4**	**16.3**	**29.5**	35.6	22.7

Table 5. Localization performance compared to the state-of-the-art on the PASCAL VOC 2008 dataset.

[12] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *PAMI*, 30(1):36–51, 2008.

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

[14] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[15] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *PAMI*, 20(3):226–239, 1998.

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

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

[18] I. Laptev. Improvements of object detection using boosted histograms. In *BMVC*, 2006.

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

[20] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

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

[23] F. Perronnin. Universal and adapted vocabularies for generic visual categorization. *PAMI*, 30(7):1243–1256, 2008.

[24] H. A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. In *NIPS*, 1995.

[25] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[26] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, 2003.

[27] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *ECCV*, 2008.

[28] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[29] L. Wolf and S. Bileschi. A critical view of context. *IJCV*, 69(2):251–261, 2006.

[30] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *IJCV*, 73(2):213–238, 2007.

[31] Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*, 2006.