Int J Comput Vis (2012) 96:46–63
DOI 10.1007/s11263-011-0443-1

The Visual Extent of an Object

Suppose We Know the Object Locations

J.R.R. Uijlings · A.W.M. Smeulders · R.J.H. Scha

Received: 27 July 2010 / Accepted: 28 March 2011 / Published online: 10 May 2011

Abstract The visual extent of an object reaches beyond the object itself. This is a long standing fact in psychology and is reflected in image retrieval techniques which aggregate statistics from the whole image in order to identify the object within. However, it is unclear to what degree and how the visual extent of an object affects classification performance. In this paper we investigate the visual extent of an object on the Pascal VOC dataset using a Bag-of-Words implementation with (colour) SIFT descriptors.

Our analysis is performed from two angles. (a) Not knowing the object location, we determine where in the image the support for object classification resides. We call this the *normal* situation. (b) Assuming that the object location is known, we evaluate the relative potential of the object and its surround, and of the object border and object interior. We call this the *ideal* situation. Our most important discoveries are: (i) Surroundings can adequately distinguish between groups of classes: furniture, animals, and land-vehicles. For distinguishing categories within one group the surroundings become a source of confusion. (ii) The physically rigid *plane, bike, bus, car,* and *train* classes are recognised by interior boundaries and shape, not by texture. The non-rigid animals *dog, cat, cow,* and *sheep* are recognised primarily by texture, i.e. fur, as their projected shape varies greatly. (iii) We confirm an early observation from human psychology (Biederman in Perceptual Organization, pp. 213–263, 1981): in the ideal situation with known object locations, recognition is no longer improved by considering surroundings. In contrast, in the normal situation with unknown object locations, the surroundings significantly contribute to the recognition of most classes.

Keywords Content based image retrieval · Visual extent · Context

Electronic supplementary material The online version of this article (doi:10.1007/s11263-011-0443-1) contains supplementary material, which is available to authorised users.

J.R.R. Uijlings (✉) · A.W.M. Smeulders
Institute for Informatics, ISIS Lab, Science Park 107, 1098 XG, Amsterdam, The Netherlands
e-mail: JRR.Uijlings@uva.nl

R.J.H. Scha
Institute for Logic, Language and Computation, Amsterdam, The Netherlands

1 Introduction

It is widely acknowledged that the visual extent of an object extends beyond the object itself (e.g. Bar 2004; Biederman 1981; Oliva and Torralba 2007; Wolf and Bileschi 2006). Nevertheless, in the early days of computer vision the visual extent of the object was sought to be precisely confined to its silhouette. And for good reasons as object boundaries (i) are more stable against lighting changes than the rest of the surface, (ii) indicate the object geometry directly, and (iii) reduce the processing requirements. This led to the idea that an object should be correctly segmented before it can be recognised. But the general task of finding the contour-bounded location of an object is very hard to solve and not really necessary for object recognition (Smeulders et al. 2000). In recent years, the use of powerful local descriptors, the increasing size of datasets to learn from, and the great advances in statistical pattern recognition have circumvented the necessity to know the object location before object-based image classification.

The first step on the road to less localisation of the object was to use local region descriptors in a specific spatial arrangement (Agarwal et al. 2004; Burl et al. 1998;

Fergus et al. 2003). This allowed the object to be found based on only its discriminative features. The second step was the introduction of the Bag-of-Words method (Sivic and Zisserman 2003), which selects interesting regions, converts them to visual words, and uses word counts followed by a spatial verification step to retrieve matching image regions. In the third step, Csurka et al. (2004) generalised Bag-of-Words to image classification and removed the spatial verification, relying on interest point detectors to extract visual words from the object. In the final step, the quantity of visual words was found to be more important than the quality of the location of the visual words (Jurie and Triggs 2005; Nowak et al. 2006). Therefore these words are no longer extracted at salient points but on a dense, regular grid. This has caused the last notion of object location to be lost in the Bag-of-Words representation which therefore mixes context and object indiscriminately. This is the state-of-the-art of image classification in 2009 (Everingham et al. 2010; Smeaton and Over 2006).

While discarding the object location has its advantages, it is also unsatisfactory. On the one hand, discarding the object location leads to computational benefits and a natural incorporation of context. On the other hand, it is unclear how much information is lost by discarding the object location: the object features of a small object in a large field of view are drowned in the information of its surroundings. Therefore this paper investigates the question: What is the visual extent of an object? This paper is an extension of Uijlings et al. (2009). Specifically, we investigate the relative influence of the object and its surroundings, and of the object interior and object border.

2 Related Work

The influence of context on recognition was researched earlier in human vision. Most notably, Biederman (Biederman 1981) considered five types of relations between the object and its context: (1) *Support* reflects that objects do not float in the air. (2) *Interposition* deals with occlusion. (3) *Probability* is the likelihood that an object is present given the context. (4) *Position* is the location within the image (e.g. a knife can be found next to a fork). And (5) *size* is the familiar size of the object. He measured the time it took for humans to identify objects violating one or more of the constraints, which reflects the difficult of identification. In this paper we focus on Biederman's *probability* by automatic rather than human vision, leaving the remaining four to another occasion. We measure the difficulty of identification in terms of classification accuracy.

Oliva and Torralba (2007) give a good overview of work in visual cognition and cognitive neuroscience on visual context and place this in light of recent advances in computer vision. They conclude that although real-world relationships between individual objects seems the most complete way to describe context, context is already described effectively by its global statistics which ignores object identities and their relations. This was also observed in earlier experimental work in computer vision by Wolf and Bileschi (2006), who showed that high-level semantic context (i.e. the co-occurrence of *buildings, trees, sky*, etc.) provided no additional information over low-level image statistics. In our paper, we represent context as a Bag-of-Words representation which can be seen as a form of low-level global image statistics.

The use of the term "context" in computer vision is rather broad. To make the terminology more precise, Divvala et al. (2009) identify several types of context as used in the computer vision community. These include Local Pixel Context (Carbonetto et al. 2004; Dalal and Triggs 2005; Fulkerson et al. 2009; Gould et al. 2009; Shotton et al. 2009), 2D scene gist context (Oliva and Torralba 2001), 3D geometric context (Hoiem et al. 2008; Nedović and Smeulders 2010), and semantic context (Malisiewicz and Efros 2009; Rabinovich et al. 2007; Singhal et al. 2003). In their definition the Local Pixel Context captures the contextual information in terms of low-level image statistics while Semantic Context captures contextual information in terms of meaningful categories (e.g. scene class or object class). In accordance with the best image retrieval methods, in this paper we study the visual extent of an object through the use of low-level features rather than semantics; we do not use region class labels as in Markov Random Fields or Conditional Random Fields (Carbonetto et al. 2004; Shotton et al. 2009) and we do not use a scene label, but we directly use the features which we extract from the image.

Zhang et al. (2007) studied the influence of context in their work. They concluded that the influence of context is marginal within the Bag-of-Words framework. However, the dataset on which they tested it consists of only four classes. On the larger and more diverse Pascal 2010 dataset, we will challenge this finding and investigate whether the influence of context in the Bag-of-Words framework is significant.

Tuytelaars and Schmid (2007) visualised a pixel-wise classification based on visual words. Using a large visual vocabulary extracted from a regular lattice, they calculated the likelihood of each visual word belonging to an object. Using an independence assumption on the visual words in the image, they used this likelihood to calculate for each pixel the probability of belonging to a certain object class. This led to an increased insight in Bag-of-Words. Similarly, in our paper we calculate for each pixel how much it contributes to the classifier output. However, as we calculate this contribution from the complete image representation rather than the individual visual words, we do not use an independence

assumption. Instead, we provide a direct visualisation of the classification of a state-of-the-art Bag-of-Words framework.

Blaschko and Lampert (2009) employed context to improve object localisation. But rather than only relying on only global context, they explicitly optimise over the amount of local context around the object. In this paper we use global context, but we investigate the influence of amount of context relative to the size of the object.

Harzallah et al. (2009) successfully combined object localisation with object classification for content based image retrieval. Their work can be interpreted as combining object features from the localised object with context features taken from the whole image. Within video, Ullah et al. (2010) automatically created object/surround distinctions using motion and object detectors, successfully improving over their normal Bag-of-Words baseline. Both works show that modelling the object location improves results. In this paper we provide an upper bound of retrieval performance when the object is localised, and compare this with the improvements obtained by Harzallah et al. (2009). Note however, that we give this upper bound while using bounding boxes. As Malisiewicz and Efros (2007) showed, this bound is even higher when the object is localised by an accurate segmentation.

3 Methodology

This paper investigates the visual extent of an object in image classification. Over the years, the Bag-of-Words method has been established as the best framework in the major retrieval benchmarks such as the TRECVID high-level feature extraction task for retrieving video (Smeaton and Over 2006) and the Pascal VOC Classification task for retrieving images (Everingham et al. 2010). In this paper we build on our state-of-the-art Bag-of-Words pipeline which won the Pascal VOC 2008 classification task and which was a runner-up in 2009.

We follow two lines in our investigation, visualised in Fig. 1. The first line is the *normal* situation where we apply a visual concept detection algorithm and determine which image parts contribute how much in identifying the target object. The second line is the *ideal* situation where we use the known object locations to isolate the object, surround, and object interior and object border. For each of these image parts we create a separate representation and examine their retrieval performance. The first line shows what currently *is* measured, and the second reveals what *could* be measured.

We investigate the visual extent of an object in the Bag-of-Words framework in terms of the object surround, object border, and object interior. We split this in two separate experiments. In one experiment we investigate the influence of the surround with respect to the complete object. In the other experiment we investigate the influence of the object border with respect to the object interior.

We use the ground truth object locations to isolate the object from its surround in both lines of our investigation. As the Bag-of-Words framework thrives using lots of data, we use a large dataset where the locations are given in terms of bounding boxes. To make a better distinction between object and surround and object interior and object border, we also perform the analysis on a smaller dataset where the locations are given in terms of a segmentation. In the normal situation we make the distinction between object/surround and interior/border after classification on the test set only. In the ideal situation we make this distinction beforehand on both the training and test set. When there are multiple instances of the same class we combine their measurements to avoid measuring object features in its surround.

3.1 Dataset

We choose to use datasets from the widely used Pascal VOC challenge as this allows for a good interpretation and comparison with respect to other work. We benchmark our Bag-of-Words algorithm on the Pascal VOC 2007 classification challenge to show our framework is competitive. Our analysis is done on two Pascal VOC 2010 datasets. First, we use the classification dataset which provides the object locations in terms of bounding boxes. In this dataset we emphasise quantity of annotations over the quality of annotations. Second, we use the segmentation dataset which is much smaller but provides more accurate object locations in terms of segments. For the Pascal VOC 2010 datasets we use the predefined `train` set for training and the `val` set for testing.

The Pascal VOC datasets we use consist images from www.flickr.com, containing twenty different object classes: *aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow*, *dining-table, dog, horse, motorbike, person, potted-plant*, *sheep, sofa, train*, and *TV/monitor*. Some images contain multiple classes. The 2010 classification set consist of 4998 `train` images and 5105 `val` images. The 2010 segmentation set consists of 964 `train` images and 964 `val` images.

Classification performance of the Pascal VOC dataset is measured by the interpolated Average Precision of a ranked list. In this paper we use the more standard Average Precision as it enables us to create a confusion matrix as we present shortly. The Average Precision is defined as

$$\mathrm{AP} = \frac{1}{m} \sum_{i=1}^{n} \frac{f_c(x_i)}{i}, \tag{1}$$

where: n is the number of images. m is the number of images of class c. x_i is the i-th image in the ranked list $X = \{x_1, \ldots, x_n\}$. Finally, f_c is a function which returns the

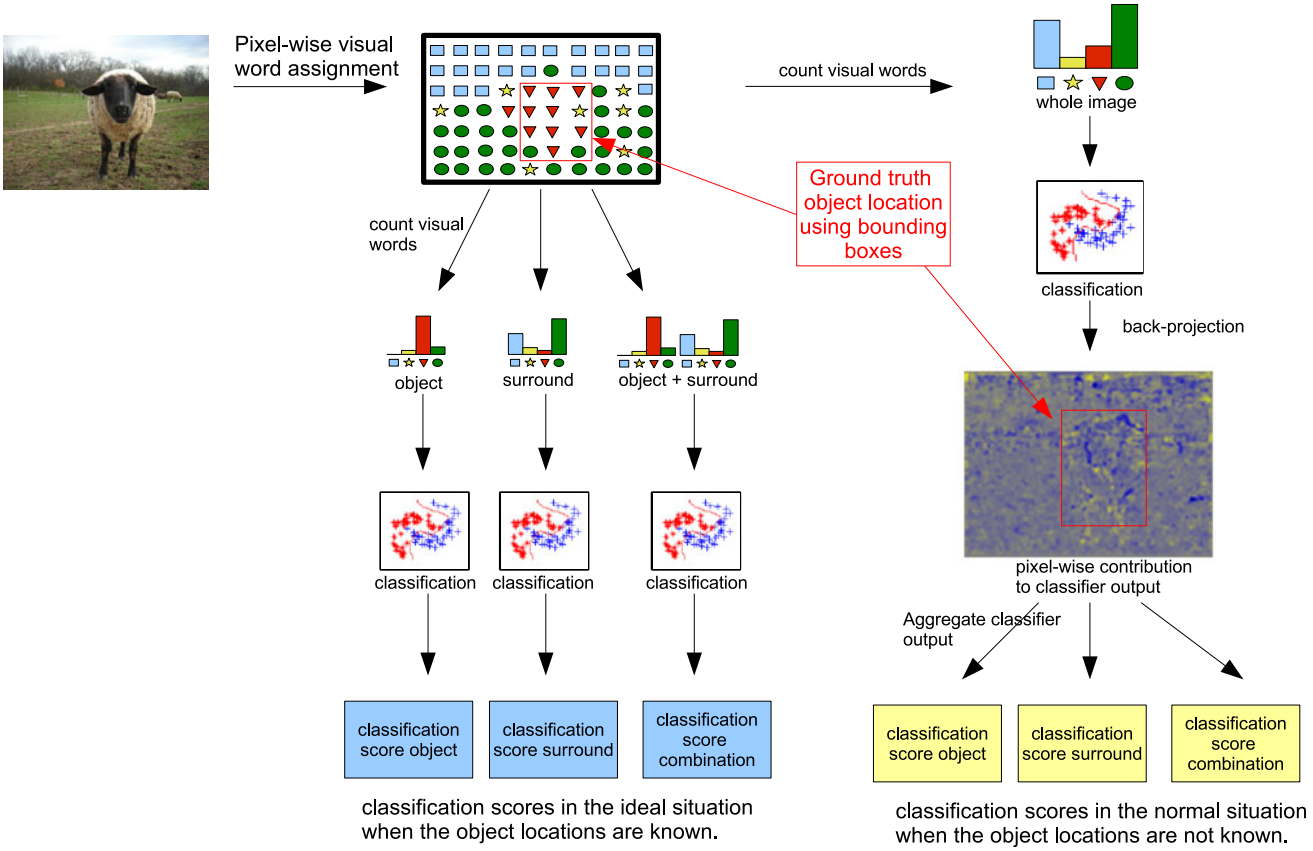


Fig. 1 The two main lines of our analysis: The *ideal line on the left* uses the ground truth object locations to divide the image into object and surround, and object interior and object border before classification. The *normal line on the right* first classifies the image, projects the classification score back on the image and then aggregates classifier scores over the object and surround, and object interior and object border

number of images of class c in the first i images if x_i is of class c and 0 otherwise. This measure has range $(0, 1]$ where a higher number means better performance.

3.2 Evaluation Matrix

To facilitate analysis, we developed a confusion matrix based on the Average Precision, which we call Confusion Average Precision Matrix or CAMP. The CAMP includes the Average Precision in its diagonal elements and, similar to a confusion matrix, shows which classes are confused.

We define the confusion or off-diagonal elements of the CAMP as the total loss of Average Precision of encountering a specific non-target class in the ranked list. To calculate the loss we traverse the ranked list in decreasing order of importance. When a non-target class is encountered at position i, the loss L is the difference between the AP assuming a perfect ranking from position i and the AP assuming a perfect ranking from position $i+1$. More formally, let $\hat{f}_c$ be a function which returns the number of examples of class c in the first i entries in the ranked list, and let $r = m - \hat{f}_c(x_i)$. Now we can calculate the loss L at position i as

$$L(x_i) = \frac{1}{m}\left(\sum_{j=1}^{r} \frac{\hat{f}_c(x_i)+j}{i+j-1} - \sum_{j=1}^{r} \frac{\hat{f}_c(x_i)+j}{i+j}\right). \tag{2}$$

The total confusion with a non-target class d is the sum of loss to that class, calculated by $\sum_{x_i \in d} L(x_i)$. As we measure confusion in terms of loss, by definition the AP plus the sum of the loss over all classes adds to one.

3.3 Bag-of-Words Framework

A condense overview of our Bag-of-Words implementation (Uijlings et al. 2010) is given in Table 1. We sample small regions at each pixel which is an extreme form of sampling using a regular, dense grid (Jurie and Triggs 2005; Nowak et al. 2006). From these regions we extract SIFT (Lowe 2004) and four colour SIFT variants (van de Sande et al. 2010) which have been shown to be superior for image retrieval (Mikolajczyk and Schmid 2005; van de Sande et al. 2010; Zhang et al. 2007). Thus we use intensity-based

SIFT, opponent-SIFT, rg-SIFT (normalised RGB), RGB-SIFT, and C-SIFT. Normally, SIFT consists of 4 by 4 subregions. However, we want our descriptors to be as small as possible in our experiments to be able to make the distinctions between object interior, object border, and object surround as crisp as possible. We therefore extract SIFT features of 2 by 2 subregions, which degrades performance no more than 0.02 MAP as shown in Sect. 4.1. The size of such SIFT patch is 8 by 8 pixels. We later verify our results on normal 4×4 SIFT, which is 16 by 16 pixels.

For the creation of a visual vocabulary we use a Random Forest (Moosmann et al. 2006) in combination with PCA on the descriptors to reduce the dimensionality by a factor 2. This yields equally accurate results as using a k-means visual vocabulary, yet is much faster (Moosmann et al. 2006; Uijlings et al. 2010). Our Random Forest consists of 4 trees of depth 10, resulting in a total size of 4,096 visual words. To train a tree from the Random Forest we use the extremely randomised trees algorithm (Geurts et al. 2006), using 500,000 labelled descriptors sampled randomly from the training set, where the labels are obtained from the annotation at image level.

For classification we use a Support Vector Machine (SVM), which is currently the most popular classifier in Bag-of-Words due to its robustness against large feature vectors and sparse data. The χ^2 kernel was found to be the best choice for the kernel function (Jiang et al. 2007; Zhang et al. 2007). However, we use the Histogram Intersection based SVM, which allows us to back-project the output of the classifier onto the image as we explain in Sect. 3.4.1. By taking the square root of the visual word histograms before normalisation we compensate for high frequent visual words, which makes the Histogram Intersection kernel almost as good as the χ^2 kernel (Uijlings et al. 2010). In fact, when we sample visual words every pixel, both the χ^2 kernel and the histogram intersection kernel yield similar accuracy.

The original Bag of Words framework is orderless. Therefore Lazebnik et al. (2006) introduced a weak spatial order by using their spatial pyramid, in which an image is divided into regular subregions. Visual word frequency histograms are obtained from each region separately. We use the spatial pyramid in half of our experiments. In the normal setting we create visual word histograms for the whole image and a subdivision into three horizontal segments, shown to be a good division by several researchers (Marszałek et al. 2007; Tahir et al. 2008; Uijlings et al. 2010). In the ideal setting we divide the image into the three subregions representing surround, object interior and object border by using the ground truth bounding boxes. To keep the total size of the final histogram representations similar we refrain from using the spatial pyramid in the ideal setting. This omission means that the upper bound of retrieval performance in the ideal setting is underestimated. It does not influence the general conclusions of this paper.

Table 1 Overview of our Bag-of-Words implementation. In our two lines of analysis we divide the image into subregions by either using the Spatial Pyramid or the ground truth object locations

Descriptor extraction	Word assignment	Classification
• Sampling each pixel	• PCA dimension reduction by 50%	• SVM:
• Size: 8×8 pixels		– Hist Int kernel
• Descriptors:	• Random Forest: 4 binary decision trees of depth 10	• Image Divisions:
– 2×2 SIFT		⋆ Spatial Pyramid
– 2×2 opp-SIFT		– 1×1, 1×3
– 2×2 rg-SIFT		⋆ Ground truth loc.
– 2×2 RGB-SIFT		– object/surround
– 2×2 C-SIFT		– interior/border

3.4 Analysis Without Knowing the Object Location

The line of analysis where the object locations are unknown shows how all parts of the image are used for classification by current state-of-the-art methods. We first classify images using a standard, state-of-the-art Bag-of-Words framework. After classification, we project the output of the classifier back onto the image to obtain a visualisation of pixel-wise classifier contributions; the sum of the pixel-wise contributions is equal to the output of the original classifier, which measures the distance to the decision boundary.

After we have created the pixel-wise classifier contributions, we use the ground truth object locations to determine how much each image part (i.e. surround, object, object interior, object border) contributes to the classification. When an image contains multiple objects of the same class, we add contributions of all its locations together. When an image contains the target class, its location is used to make the distinction into object, surround, object interior, and object border. If the image does not contain the target class, we use the class with the highest classification contribution to make this distinction. This allows us to create a partitioning for both target and non-target images, which we need in order to calculate the Average Precision that is defined over the whole dataset (there is no "true" partitioning into the object and its surround for non-target images).

3.4.1 Back-projection of the Classifier Score

We want to determine the relative contribution of each pixel in the image. This requires dissecting the classification function to determine the relative contribution of each visual word in the image. We follow Maji et al. (2008) to rewrite the Histogram Intersection kernel, but in principle any additive kernel can be used (Vedaldi and Zisserman 2010).

The classification function for a Support Vector Machine can be written as (Bishop 2006)

$$h(\mathbf{x}) = b + \sum_{j=1}^{m} \alpha_j t_j k(\mathbf{x}, \mathbf{z}_j), \tag{3}$$

where $\mathbf{x} = \{x_1, \ldots, x_n\}$ is the vector to be classified, $\mathbf{z}_j = \{z_{1j}, \ldots, z_{nj}\}$ is the j-th support vector, α_j is its corresponding positive weight, $t_j \in \{-1, 1\}$ is its corresponding label, m is the number of support vectors, and $k(\cdot, \cdot)$ is a kernel function. For the Histogram Intersection kernel

$$k(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \min(x_i, z_i), \tag{4}$$

the classification function can be written as (Maji et al. 2008)

$$\begin{aligned} h(\mathbf{x}) &= b + \sum_{j=1}^{m} \alpha_j t_j k(\mathbf{x}, \mathbf{z}_j) \\ &= b + \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_j t_j \min(x_i, z_{ij}). \end{aligned} \tag{5}$$

As the outer sum in (5) is over the visual words, the contribution per visual word channel w_i is calculated as

$$w_i = \sum_{j=1}^{m} \alpha_j t_j \min(x_i, z_{ij}). \tag{6}$$

Within an image there are often multiple visual words having the same identity i. We evenly distribute the contribution w_i over all visual words with identity i. This gives us per visual word in the image its contribution to the classifier score. Using the locations of the patches which generated the visual words, we can project these contributions back onto the image. Examples are shown in Fig. 3.

3.5 Analysis Using the Ideal Object Location

In this line of analysis we use the known object locations to create different representations of the surround, object, object interior, and object border in both the training and test set, yielding hypothetical classification scores. We assign descriptors to an image part based on its centre point. For example, a descriptor is considered to come from an object when its centre is contained within the bounding box of that object. We use the ground truth object locations to create a separate visual word histogram for each of the image parts and analyse their retrieval performance. We create combinations by concatenating these word histograms.

Again, if an image contain multiple objects of the same class we combine their visual word histograms by adding them together. If the image contains the target class its location is used to divide the image into object, surround, object interior, and object border. If the image does not contain the target class, the class with the highest classification score is used to make this distinction. Note that if we would only use the locations and not the labels, i.e. we would always select the class with the highest classification score regardless if that is the target class or not, accuracy could only improve: in the rare cases that a non-target class has a higher classification score than the target class in that same image, the positive image will have a higher ranking. Hence the scores presented in this paper for the ideal setting can be seen as an upper bound if the locations of the objects are known.

3.6 Distinguishing Object, Surround, Interior, and Border

For boxes, the ground truth locations separate the object from the surround. Note that the nature of the boxes cause some surround to be contained in the object. To separate the object interior from the object border, we define an object interior box as being a factor n smaller than the complete object box while its centre pixel remains the same. To determine the interior box we use the idea that object border contains the shape and the object interior contains texture and interior boundaries, which should be complementary. Separating complementary information should yield better results for the combination, hence we find the optimal interior boxes by optimising classification accuracy over n on the training set using cross-validation. We found a factor 0.7 to be optimal. This means that 49% of the object is interior and the rest border.

For segments the Pascal VOC dataset only annotates the interior of the object while there is a 5 pixel zone around where the borders of the objects are. We want to ensure that no surround descriptors measure this border zone, and no interior descriptors measure this border zone. As we use the middle of our descriptor as point of reference in the ideal situation, we extend this border zone with half our descriptor size both inwards and outwards. Extending the border outwards yields our outlines of the object. Extending the border inwards yields the separation between object interior and object border. Our object border hence becomes 13 pixels wide. We measured that on average over all objects, 46% of the object becomes interior and the rest border.

4 Results

4.1 Classification Without Knowing the Object Location

We first benchmark our Bag-of-Words system on the Pascal VOC 2007 dataset, on which most results are published. For our normal Bag-of-Words system where we do

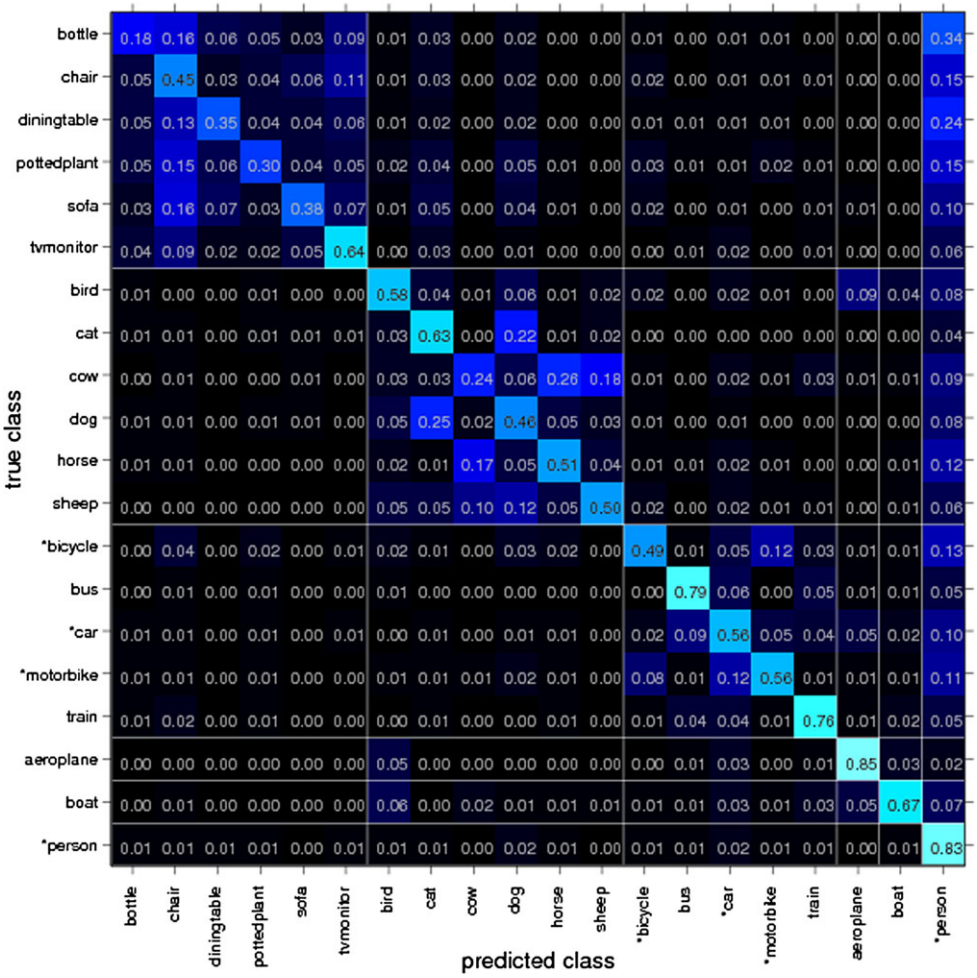


Fig. 2 Average Precision Confusion Matrix (CAMP) of the normal situation where the object locations are unknown (using 2×2 SIFT). By definition of the CAMP the rows sum to one (see Sect. 3.2). Notice the within-category confusion in the furniture, animal, and land-vehicle classes

not know the object location we achieve an accuracy of 0.57 MAP, sufficiently close to recent state-of-the-art Bag-of-Word scores obtained by Harzallah et al. (2009) and van de Sande et al. (2010), which are respectively 0.60 MAP and 0.61 MAP. To enable back-projection with (6) we use the Histogram Intersection kernel instead of the widely accepted χ^2 kernel (Harzallah et al. 2009; Jiang et al. 2007; van de Sande et al. 2010; Zhang et al. 2007). This does not influence classification accuracy: with the χ^2 kernel performance stays at 0.57 MAP. Instead, most of the difference in accuracy between our work and Harzallah et al. (2009), van de Sande et al. (2010) can be attributed to our use of 2×2 SIFT patches: using the four times as large 4×4 SIFT descriptor results in a classification accuracy of 0.59 MAP. However, in most of our experiments we favour small SIFT descriptors to minimise the overlap between object and surround, and interior and border descriptors. From now on all results are reported on the Pascal VOC 2010 dataset using 2×2 SIFT descriptors, unless otherwise noted.

Figure 2 shows the confusion matrix of the normal Bag-of-Words system on the 2010 `train+val` set. One can see that the classes can be roughly divided into three clusters where most of the confusion concentrates: *furniture*, *animals*, and *land-vehicles*. The classes *aeroplane*, *boat*, and *person* behave differently and cannot be grouped. The high confusion with the *person* class in the right column of Fig. 2 can be explained by the many *person* images in the dataset. We will use the identified categories in subsequent analysis.

To conclude, we have verified that our Bag-of-Words system is state of the art and we have identified categories to facilitate subsequent analysis.

4.1.1 *Localising the Classifier Contributions*

We now investigate qualitatively where the Bag-of-Words classifier obtains the evidence to classify images. We do this for both 2×2 SIFT used in most of our paper and the more widely used 4×4 SIFT. We use the method described in Sect. 3.4.1 and show results for top-ranked images (according to 2×2 SIFT) of the classes *aeroplane, boat, cat, car, person*, and *sofa* in Fig. 3.

We first observe that the difference between the use of 4×4 and 2×2 SIFT descriptors is very small. The former seems to be a blurred version of the latter. Hence the following observations hold for both types of descriptors.

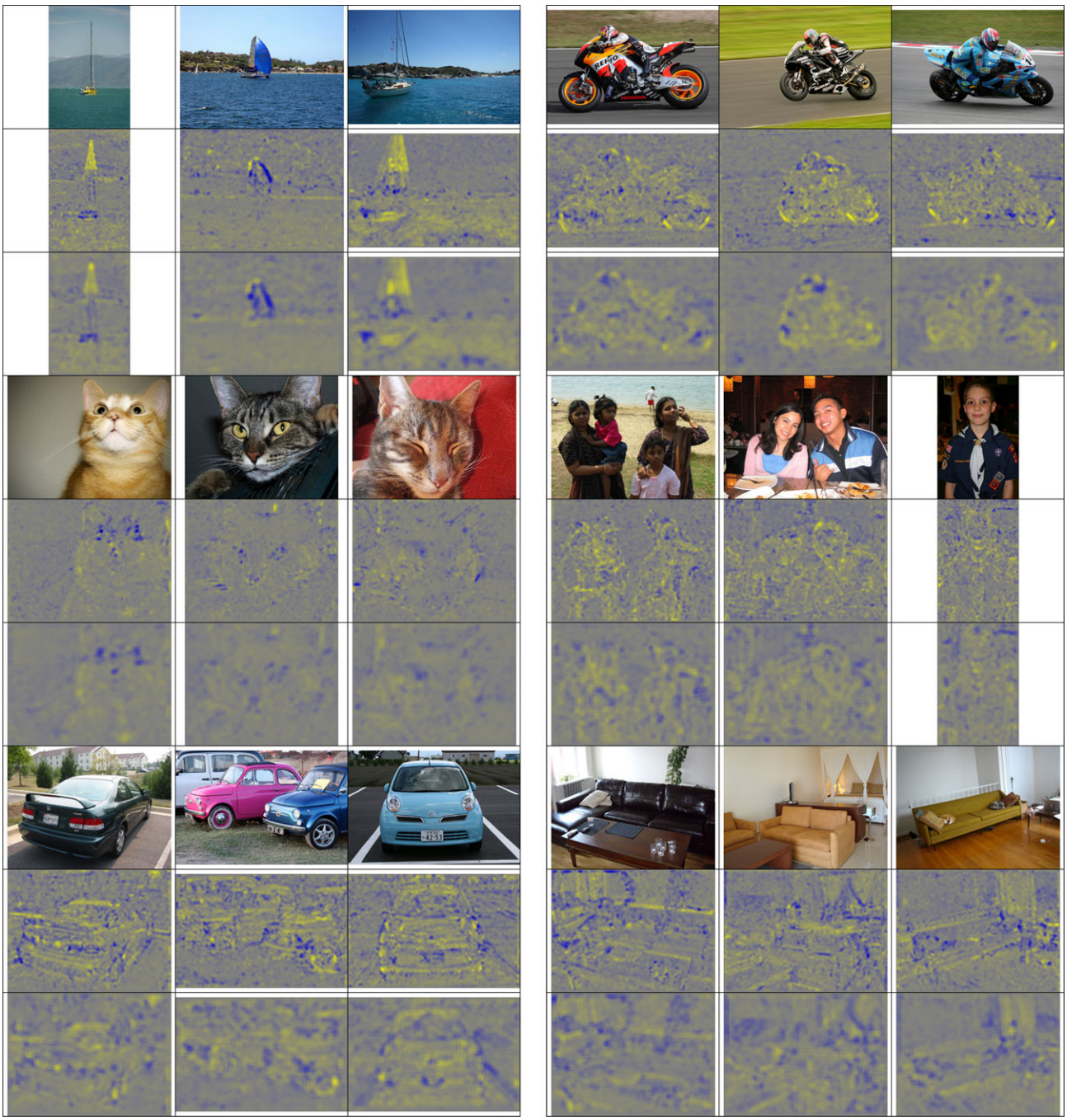


Fig. 3 Pixel-wise contribution to the classification for top ranked images for the categories *boat, cat, car, motorbike, person*, and *sofa*. The original image is followed by the contribution of 2×2 and 4×4 SIFT respectively. *Dark-blue* means a negative and *light-yellow* means a positive contribution to the classifier. Notice that high positive or high negative contributions are often located on small details. The 4×4 SIFT images resemble a blurred version of their 2×2 counterparts

We can see that, generally, in the Bag-of-Words method often small details give either a high positive or high negative contribution to the classifier output. However, while details often stretch beyond the size of the descriptor patch, as seen for example in the ropes of the boats or the contours of the cars and persons, they never coherently cover a complete object or object part. The contours of the cars come closest, but these contours are frequently interrupted by small details with a strong negative response. In homogeneous regions the responses show a considerable amount of noise, as seen for example in the erratic responses of the sky in the *boat* images. This is possibly caused by local normalisation of the

descriptors. Of course, the Bag-of-Words method was designed to work on local details but these visualisations show just how fragmented these details are.

For the *boat* class, water and sky yield both strong positive and negative contributions with an overall positive contribution. The horizon line, often sky-water, consistently yields positive information. This shows why sky and water are good contextual indicators of *boat*. Within the *boat* only the ropes and masts have a positive response, while their hulls have a strong negative response. In fact, the overall contribution within the *boat* region is negative(!). This shows that a *boat* is recognised only by the water and is therefore purely recognised by its function (being in the water).

For the *cat* images the fur is most discriminative. But like the sky, fur consists of a mix of positive and negative contributions which has a net positive contribution. This suggests that for these kinds of textures looking at small image patches is suboptimal. Furthermore the shape of the cat is not important. We see similar behaviour for the other animals, but for *horse* the shape of the legs are also important. This suggests that most animals are recognised based on texture rather than shape.

For *car*, the largest positive contribution to the classifier score is concentrated on the contours and interior boundaries. For the contours especially the roof of the car, the nose, and the wheels yield high positive information. For the interior boundaries the positive information often is concentrated on the lights, grill, bumper, and window-hood boundary. The importance of the contours suggests that *cars* are mainly recognised through their shape and interior boundaries.

In the *motorbike* images, all parts of the motorbike give an equal amount of positive information to the classifier score. Only the front wheel gives generally a strong positive contribution. The highest ranked negative examples of *motorbike* suggest that the strong response of its front wheel causes the confusion with the *bicycle* class shown in Fig. 2.

For *person* both its contours and inner boundaries are important. The shoulders, upper sides of the head, and the collar/neck boundary often yield a strong positive contribution. The clothes are mildly but erratically positive, yet their overall response is large because of the size of their surface.

In the *sofa* images primarily true vertical and tilted horizontal edges are important, which may be caused by a *sofa* or more likely a whole living room in perspective.

4.2 Classification in Ideal Setting with Known Object Location

In this experiment we use the object location to create a separate representation for the surrounding and the object, where the representation of the object may be split into the interior and the border of the object. We compare this with the results of normal situation where the object location is not known.

Figure 4 compares the performance of the normal situation in which the object location is not known with the ideal situation where the object location is known. Clearly, for all classes knowledge of the object location greatly increases performance. The overall accuracy of the normal situation is 0.54 MAP, the accuracy of the ideal situation when making the distinction between object and surround is 0.68 MAP (where no Spatial Pyramid is applied to the object). When creating separate representations for the surround, object interior, and object border performance increases to 0.73 MAP. This shows that the potential gain of knowing the object locations is 0.19 MAP in this dataset.

Similarly, on the segmentation dataset, in the normal situation where the object location is not known the classification accuracy is 0.44 MAP. When separating the object from the surround accuracy rises to 0.62 MAP. If we make a separation between surround, object interior, and object border accuracy improves to 0.69 MAP.

The huge difference between the accuracy without and without knowing the object location shows that the classifier cannot distinguish if visual words belong to the object or surround. We investigate the cause by determining for each visual word the probability that it occurs in an object (i.e. in any of the specified object classes), which is visualised in Fig. 5. This graph shows that 1% of the words have a larger than 90% probability of describing background. We found that these words describe mostly homogeneous texture (e.g. *sky*). In contrast, no single word has a larger than 90% probability of occurring on an object and less than 2% of the visual words occur on an object more than 75% of the cases. Note that these numbers are the same when using 4×4 SIFT. This means that no visual words exclusively describes objects and that these visual words are less specific than generally thought.

Results in this section suggest that performance for Bag-of-Words could be improved when the object location is explicitly modelled. Indeed, the work of Harzallah et al. (2009) combined a system for object detection, i.e. localising and classifying objects within images, and a Bag-of-Words object classification system by fusing their respective classifier outputs. In effect the object detection system explicitly modelled the object location on which it based its classification. The combination allowed them to successfully improve the classification score by 0.04 MAP to 0.64 MAP on the Pascal VOC 2007 dataset. However, our experiments on the Pascal VOC 2007 dataset result in an improvement of 0.20 MAP yielding an upper bound of 0.77 MAP if the object locations are known. This bound also applies if the labels are not used to select the object location as explained in Sect. 3.5. This means that while the work of Harzallah et al. (2009) is encouraging, there is still a lot of room for improvement in

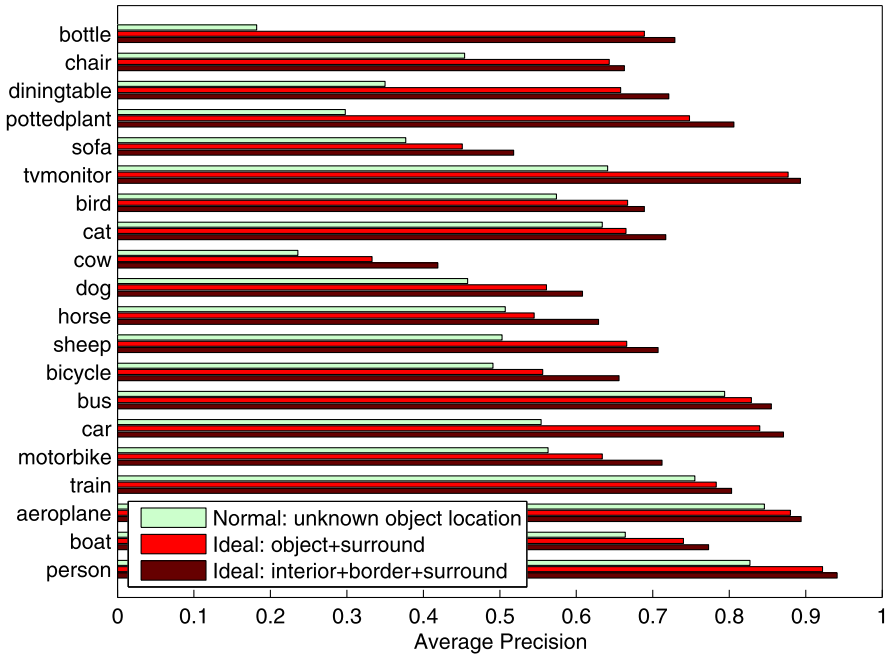


Fig. 4 A comparison of the normal situation when the object location is unknown and the ideal situation where the object location is known. Accuracy over all classes for the normal situation is 0.54 MAP, for object+surround this is 0.68 MAP, and for interior+border+surround this is 0.73 MAP

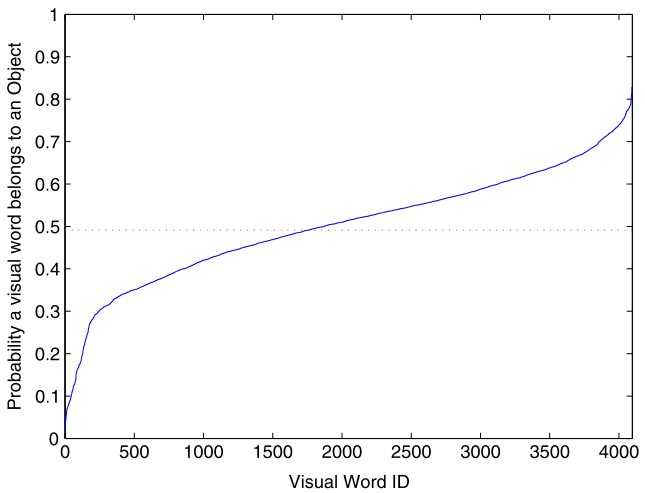


Fig. 5 The probability of each visual word belonging to an object. The *dotted red line* is the prior probability. Contrary to general belief, visual words are not very object specific as only 2% of the visual words have a higher than 75% probability to come from an object

the Bag-of-Words classification framework by attempting to locate the object within the image.

4.3 Discussion on Object Versus Surround

We now proceed to discuss the relative influence of the object and its surroundings. We do this first using boxes on the large Pascal VOC classification set. Then we perform the same on segments using the smaller Pascal VOC segmentation set.

4.3.1 Object Versus Surround Using Boxes

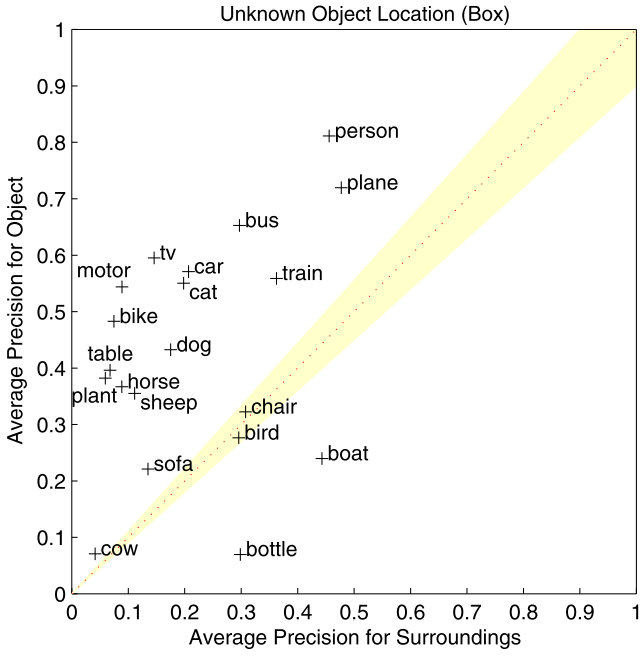
Figure 6 plots the Average Precision for the object against the surround and against the combination of the object and surround for the normal situation where the object location is unknown, Fig. 7 plots the same for the ideal setting where the object location is known.

In Fig. 6(a) one can see that for *boat* and *bottle* the surroundings are more used than the object for classification in the normal situation. For *boat* this confirms that it is recognised by only water and sky as seen in Fig. 3.

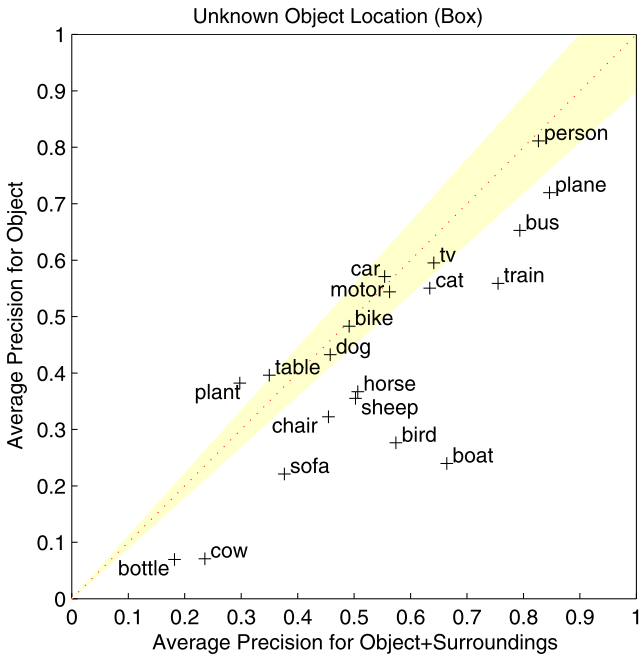
The retrieval performance when using only the surround is low for more than half of the classes in the normal setting. Only *bus, boat, bottle, bird, chair, train*, and *plane* yield reasonable performance. The performance for *person* looks also reasonable, but is close to its random score of 0.37 AP. In contrast, when training and learning on the isolated surroundings, Fig. 7(a) shows that many classes can be retrieved a lot better. Thus, while the surroundings contain information, it is normally not the focus of the classifier.

In Fig. 6(b) we see that the combination of object and surround is better than using the object alone for more than half of the classes. This is not surprising as the classifier was learned on the combination. However, for the classes *plant, table, bike, car, motor, dog, person*, and *tv/monitor* the performance of the combination is equal to using only the object. For these classes the classifier learns to ignore the surround.

When the objects are considered localised in Fig. 7(b), for all classes except *bird* and *table*, using surroundings in addition to the object does not yield much improvement over

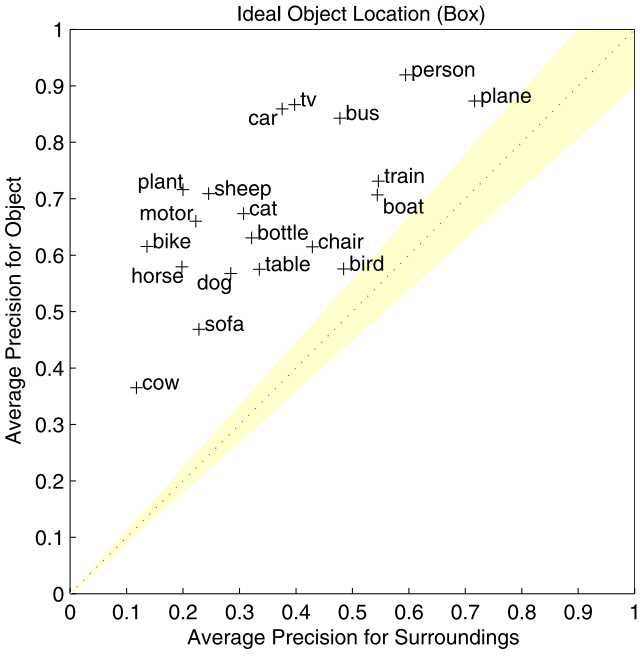


(a)

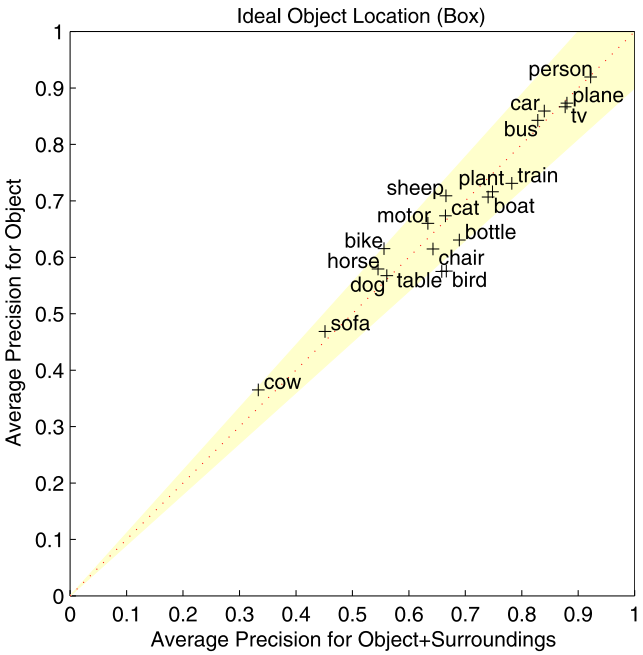


(b)

Fig. 6 The retrieval performance of the object, surround, and its combination in the normal setting where the object locations are unknown. (**a**) The surround versus the object. For *bottle* and *boat* the surround is more important than the object itself. (**b**) The object versus the combination of object and surround. For *bike, car, dog, motor, plant, person, table,* and *tv/monitor* the performance of the object is very similar to the combination, suggesting context is mostly ignored by the classifier



(a)



(b)

Fig. 7 The retrieval performance of the object, surround, and its combination in the ideal setting with known object locations. (**a**) The surround versus the object. For all classes the object is more important than the surround. For all classes performance increases significantly over Fig. 6(a). (**b**) The object versus the combination of object and surround. For most classes performance is similar for the object and the combination. Hence if the object location is considered known, the surround adds little information

using the object alone. Interestingly, this agrees with the research on human vision of Biederman (1981), who found that objects viewed in isolation are recognised equally well as objects viewed in proper context.

Intuitively, the relative size of the object and its surround will impact the results. To see how, we analysed results on two subsets of the images: one where 5–20% of the image

is object, and one where 5–20% of the image consists of the surround. In the normal situation, when the images consist mostly of object, only the *bus, boat, plane*, and *train* class can be still reasonably recognised by their surround. For the classes *bottle, bird*, and *chair* more surround is needed to adequately recognise them. For the images with large objects, adding the surrounding yields no performance improvements over using the object alone in the normal situation. When the objects within the images are small, performance drops using only the object features. However, many classes can still better be recognised by the object features: *bike, bus, car, motor, plane, cat, dog, horse, plant, table, tv,* and *person* are all better recognised by their object features than their surround. Except for *horse, plane, bus*, using the now large surround in addition to the object still does not improve recognition performance.

In the ideal setting, for both the sets with large and small surround, the surround does not add any extra information to using the object alone, except again for the *bird* and *table*. For the set with a large surround, the relative performance of the object and surround is similar to Fig. 7(a). When using only the surround for classification, recognition for images with a large surround is on average 0.11 AP higher than on the images with a small surround. In contrast, using only the object for classification, recognition for images with a large object is on average only 0.03 AP higher than for the images with a small object region, where most benefits are for the classes *bird* (0.19 AP), *chair* (0.42 AP), and *table* (0.16 AP). We conclude that the size of the surround matters in both the normal and ideal situation. For objects its size only matters in the normal situation: once the object is localised, for most objects a larger size does not result in better recognition.

We now continue with analysing the confusion matrices of using only object or surround in the idealised setting when the objects location is known, which are visualised in Fig. 8. The confusion matrix of using only surround in Fig. 8(a) looks similar to the confusion matrix of the normal setting in Fig. 2. Again, most of the confusion is concentrated within the *furniture*, *animals*, and *land-vehicle* categories. This means that each category shares context, which obviously is the case. For the *car* class something interesting happens. One can see that the *car* context is strongly confused as context for other classes, but not vice versa. This suggests that while the contexts of *bicycle*, *bus*, and *motorbike* are disjunct, the car context includes them all. Indeed, in this dataset the motorbike context is dominated by the countryside and the bus-context is dominated by urban environments, whereas the *car* occurs in both.

Figure 8(b) displays the confusion matrix when only *object* descriptors are used. Most notably, the confusion within the *furniture* and *land-vehicle* category is very low, which means that confusion within these two categories is mainly caused by the surroundings. Although without the surround bicycles continue to be confused with motorbikes, and buses with trains. For *animals*, within category confusion is still high. This means that both context and object are a source of confusion. Intuitively, object descriptors cause confusion because most of the animals are furry and have similar shapes (four legs and a head). In Sect. 4.4 we will see what causes most confusion: fur or shape.

4.3.2 Object Versus Surround Using Segments

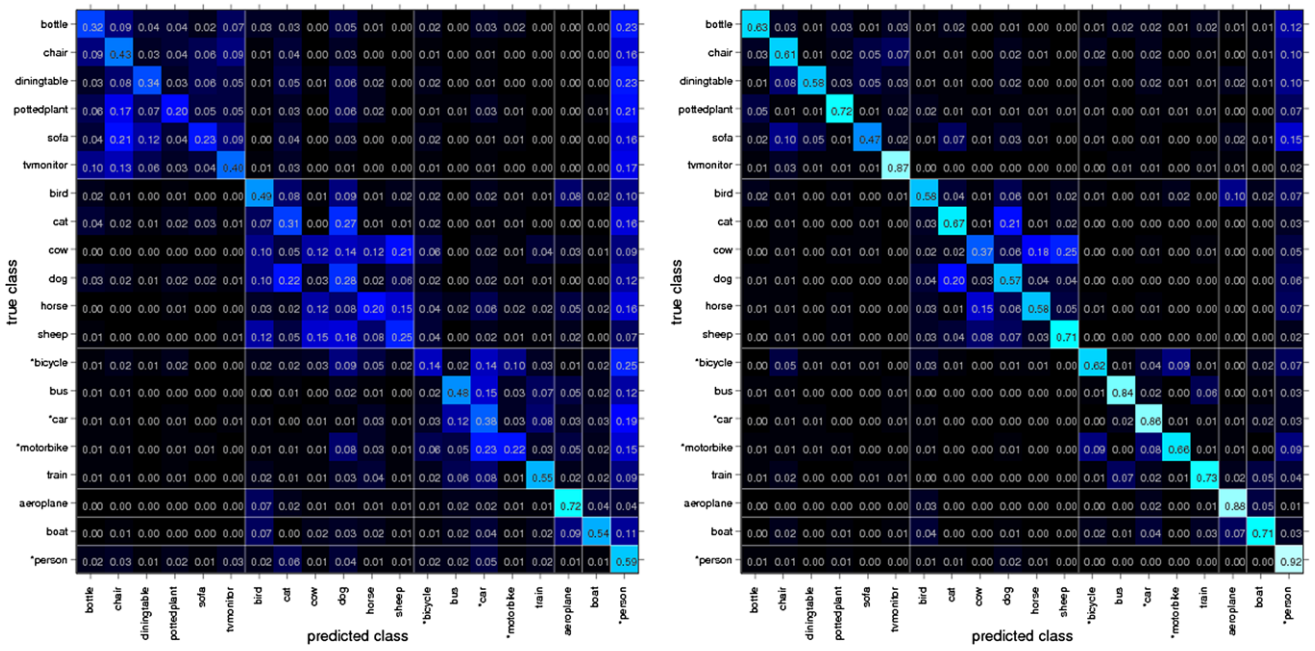
We repeated the experiments to analyse the influence of the object and the surround, but this time on fewer data but using more accurate object locations in terms of segments.

The comparison of the influence between the object and its surround in the normal situation for segments looks similar to Fig. 6(a), except that performance of using only the object is worse. Hence with fewer training examples the classifier is still able to learn the appearance of the surrounding but has less success in learning the appearance of the object itself. This means that the appearance of the context is simpler than that of an object. In effect, this also means that *cow, sofa, bird*, and *chair* join the *boat* and *bottle* class in that their surroundings are more important than the object itself when using fewer training examples.

Figure 9(a) shows that the combination of object and surround is better than using the object alone for more than half of the classes, similar as with the larger dataset on the boxes (Fig. 6(b)). Again, for *motor, dog*, and *plant* adding the surroundings does not help. For *bike*, and *tv* using also the surroundings has even a negative effect.

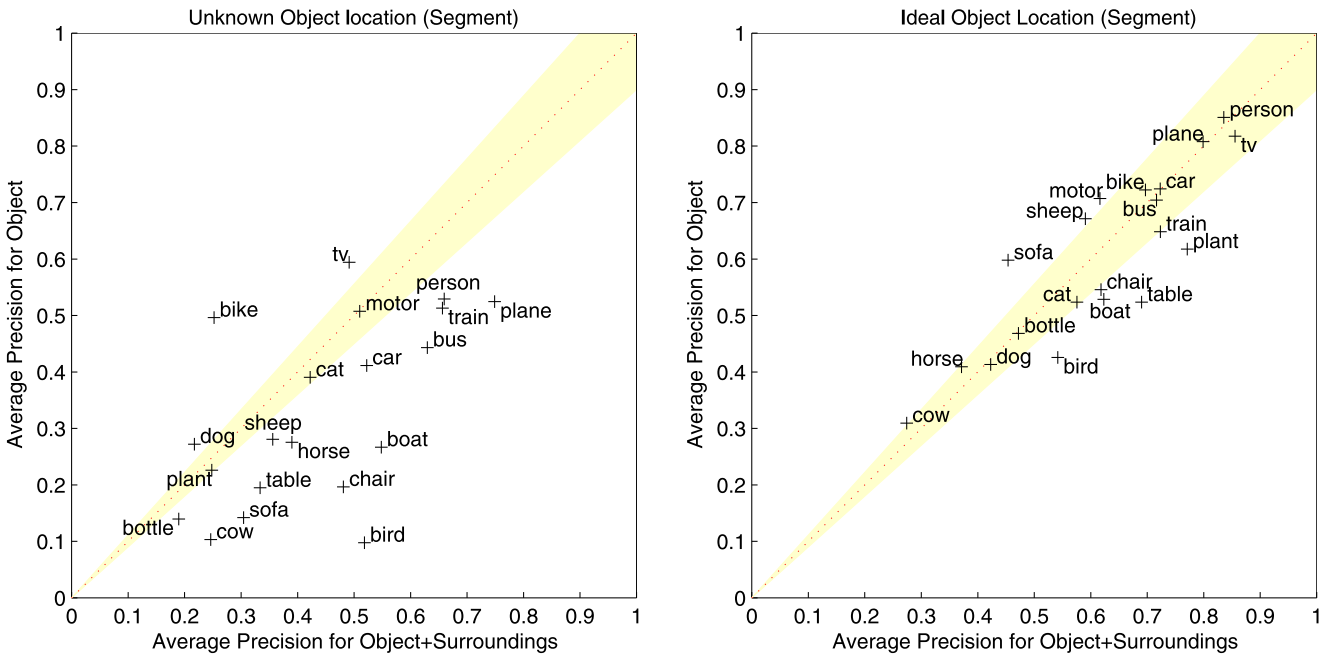
In the ideal situation, when comparing object and surround again the results are similar to Fig. 7(b). Again, most classes can be retrieved reasonably by their surround. However, in contrast to Fig. 7(a), *boat* and *bird* can be recognised equally well by their object as by their surround when this distinction is made more accurate by a segmentation. This suggests that part of the classification performance for using only object using a bounding box can be attributed to the inclusion of a bit of context.

Figure 9(b) compares the accuracy of the segmented object with the combination of the segmented object and its surround. We see that now beside *bird*, also the classes *boat, chair, plant, table* and *train* benefit from the inclusion of the surround. These classes all have a high variability in appearance, and are difficult to recognise in isolation. To verify whether this change in behaviour comes from the omission of any context while using segments, we repeated the experiment on the segmentation dataset but using boxes. Results were the same. Hence we conclude that the behaviour results from using fewer training examples: to accurately learn the appearance of these relatively difficult classes more training data is needed.



(a) (b)

Fig. 8 Confusion matrices when using only the descriptors from the surroundings and when using only the descriptors from the object. (**a**) Surround descriptors only. (**b**) Object descriptors only. Using only surround descriptors in (**a**) there is a lot of confusion within the furniture, animal, and vehicle category. These categories therefore share context. In contrast, when using only object descriptors there is only a significant confusion within the animal category. This shows that animals share many object features, but furniture and vehicles do not



(a) (b)

Fig. 9 Influence of the object and its surround analysed using segments. Results are similar to Figs. 6(b) and 7(b)

4.3.3 Conclusion Object Versus Surround

In the normal situation where the object location is unknown the surroundings contribute significantly to classification for more than half of the classes. For the classes *boat* and *bird* the surroundings are even more important then the object itself. This means that the findings of Zhang et al. (2007) that the Bag-of-Words framework can learn to ignore the surroundings holds for some classes, but does not generalise to all classes in larger datasets. In contrast, in the ideal setting when the object locations are known, the surroundings add little additional information for most classes which is in accordance with human vision (Biederman 1981). Finally, the surroundings are a source of confusion within the *furniture*, *animal*, and *land-vehicle* categories, but the object itself only causes confusion within *animals*.

4.4 Discussion on Interior Versus Border

We now discuss the relative influence of the interior and the border of the object. As the segmentation yields a more accurate distinction than the boxes, we will first discuss the results on the segmented object locations. Afterwards we will verify the observations on the larger dataset using boxes.

4.4.1 Interior Versus Border Using Segments

Figure 10 plots the Average Precision for the interior against the border and against the combination of the interior and border for the normal situation with unknown object locations. Figure 11 plots the same for the ideal situation with known object locations.

First we look at the animal classes *cat, cow, dog, horse,* and *sheep*. In both the normal and ideal situation, we see from Figs. 11(a) and 10(a) that the interior contains significantly more information than the border. In Figs. 11(b) and 10(b) we can see that adding the border as additional information does not improve results, except for the *horse* when the object location is known. We conclude that the animals are recognised not based on their contours but on their interior. Hence the animals are recognised based mostly on their fur, which was observed earlier for *cat* in Fig. 3.

We now consider the vehicles *car, bus,* and *train*. In Figs. 11(a) and 10(a) we see that also for these classes the interior is more important than the border, yet in contrast with the animals, the border alone still yields good accuracy. As seen in Fig. 11(b), when the object location is unknown, using the border and the interior yields little improvements over using the interior alone. But Fig. 10(b) shows that when the object location is known, using the border in addition to the interior yields improvements of around 10% for *car, bus,* and *train*. Hence both the border (shape) and interior for these classes are important, where the visualisations of the classifier contributions in Fig. 3 suggest that the interior is important because of their well-defined interior boundaries.

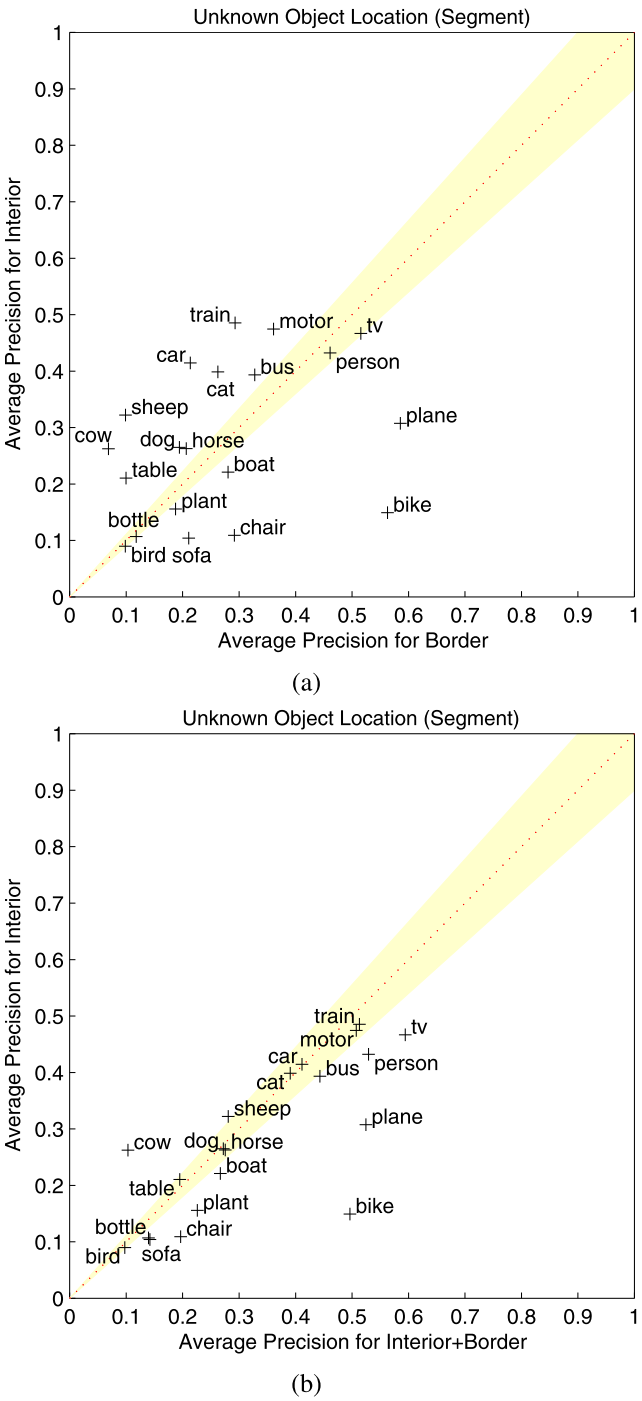
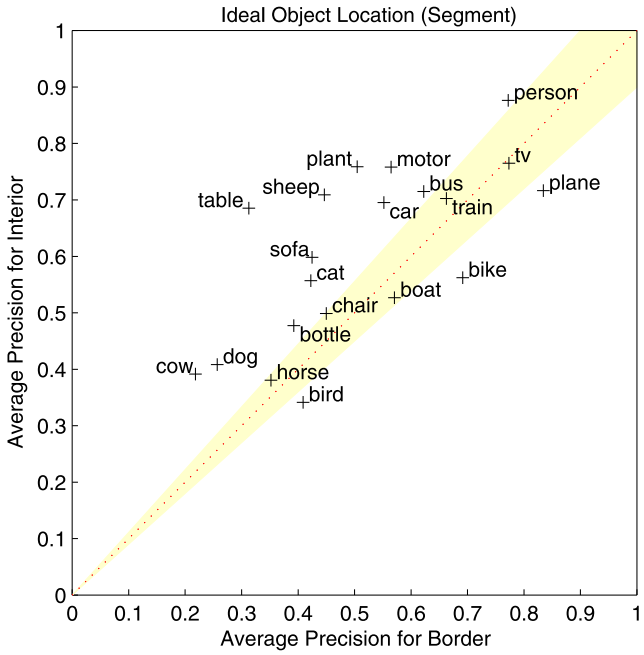
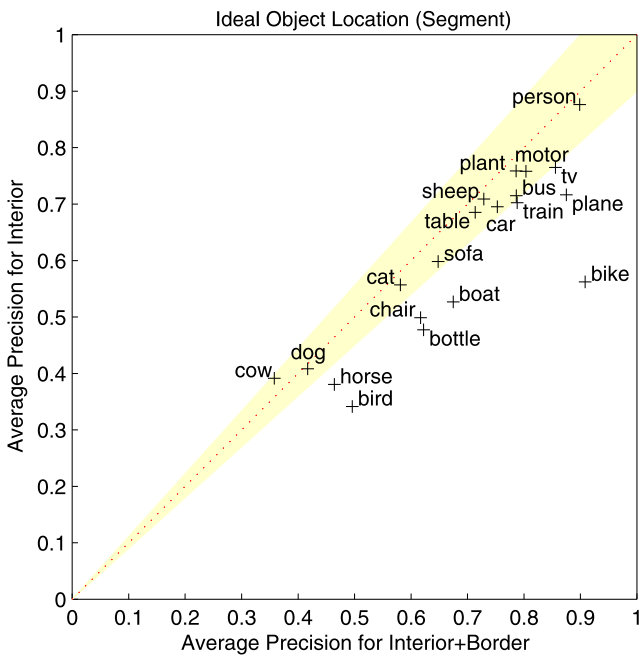


Fig. 10 The retrieval performance of the object interior, object border, and their combination in the normal situation with unknown object locations. (**a**) Object border versus object interior. The animal classes *dog, cat, cow, sheep,* and *horse* are best recognised by their interior. (**b**) Object interior versus the combination. Performance of the animal classes does not improve while using the border in addition to the interior



(a)



(b)

Fig. 11 The retrieval performance of the interior, border, and their combination in the idealised setting where the object locations are considered known. (**a**) Object border versus object interior. Most animal classes are best recognised by their interior. Most vehicle classes as well yet these are also recognised well by only their border. (**b**) Object interior versus the combination. For the animal classes, using the border in addition to the interior does not yield additional information. For the vehicle classes, the combination of the interior and border do yield improvements

For *bike* and *plane*, Figs. 11(a) and 10(a) show that the border is more important than their interior. In fact, *plane* is the only class where, when the object location is known, using the interior in conjunction with the border yields no improvements over using the border alone. For *bike* both the interior and border are important when the location is known. Hence *bike* and *plane* are rigid classes with a well-defined shape which can be recognised best by their border, while for *bike* the interior is also important.

The classes *chair, plant*, and *sofa* are the only classes that behave different in the normal and ideal situation. When the object location is unknown, Fig. 11(a) shows that the border yields more information than the interior. In contrast, for these classes when the location is known the interior yields more information in Fig. 10(a). This suggest that while the interior yields enough information to discriminate between classes, it yields not enough information to discriminate between the class and the background which is necessary in the normal situation when the object location is unknown. Indeed, intuitively, *plant* resembles any background vegetation and *sofa* may resemble carpet of curtains in the background.

4.4.2 Interior Versus Border Using Boxes

Figures 12 and 13 show the performance of using only the interior versus the performance of using only the surround when this distinction is made using bounding boxes. The tendencies are similar as in the situation where the segmentation is used to make the distinction. Again, for the animal classes *cat, dog, horse,* and *sheep* the interior is more important than the border. However, the border now has a higher performance because for the large, more easily recognisable objects it includes more of the object interior. For the *bus, car, train,* and *plane* classes both the interior and the border are equally good for predicting the object class, which corresponds with our earlier observation that these rigid classes can be recognised by their well-defined shape as well as their interior borders.

For *bike* the interior is now more important. This is because the segmentation accurately outlines the wire-frame leaving little surface for the interior. The fact that the interior is important using boxes just shows the importance of the inner frame and parts of the spokes.

4.4.3 Discussion and Conclusion Interior Versus Border

The object interior consists of texture and of interior boundaries, reasonably captured by a Bag-of-Words representation. However, this representation may be less appropriate for the object boundary as the object shape is intuitively better represented by larger fragments with more geometric constraints. However, we saw from Fig. 3 that this representation still highlights large parts of object boundaries

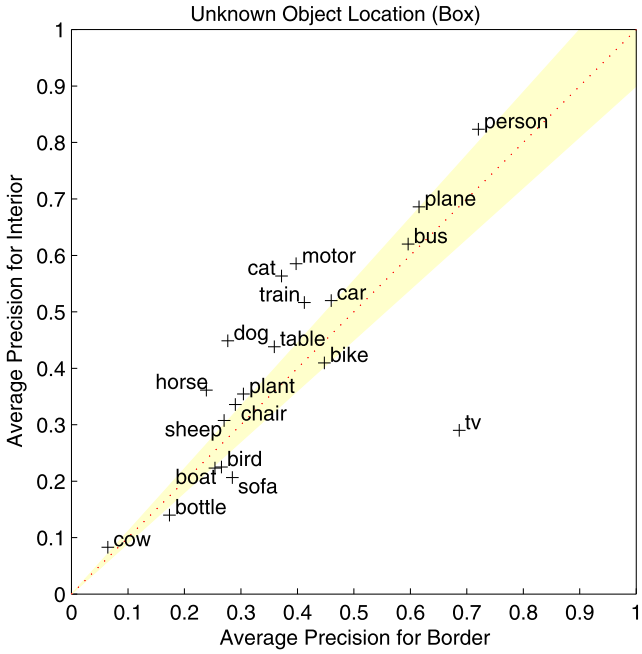


Fig. 12 The retrieval performance of the object interior, object border, and their combination for the normal situation with unknown object locations. The distinction is made through boxes. Similar to using segments (Fig. 10), the animal classes *dog, cat, horse,* and *sheep* are better recognised by their interior

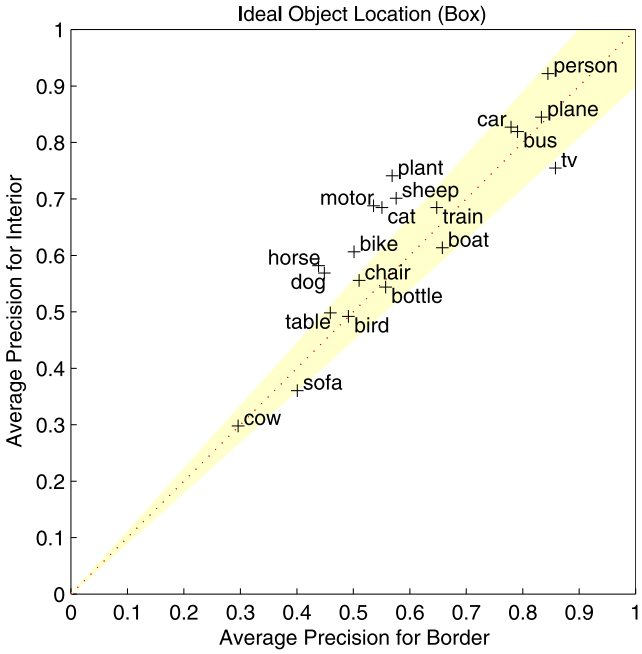


Fig. 13 The retrieval performance of the interior, border, and their combination in the idealised setting where the object locations are considered known. The distinction is made in terms of boxes. The animal classes *dog*, *cat*, *horse*, and *sheep* are best recognised by their interior. The vehicle classes *car, bus,* and *plane* are recognised equally well by the interior as their border

extending beyond the size of the local patch. Hence while the conclusions made on the relative contribution of the border and the interior may not extend to object recognition in general, it will still be indicative of the relative difficulty of obtaining information of the object border and object interior.

To conclude, our analysis of the object border and object interior showed that the non-rigid animal classes *cat, cow, dog, horse,* and *sheep* are mostly recognised by their fur while their shape adds little information. The exception is the *horse* whose legs likely contribute. For the rigid classes *bike, bus, car,* and *train*, both interior boundaries and the border or shape information is used for recognition. For *plane* only the shape is sufficient for recognition.

4.5 Using 4×4 SIFT

The results presented in this paper were based mostly on a 2×2 SIFT descriptor, as this small descriptor enabled a more crisp separation of the different image parts, especially for the interior/border distinction. To investigate the influence of this choice, we repeated some experiments using the larger 4×4 SIFT. For the object/surround distinction we repeated the experiment where the location is given by boxes. For the interior/border distinction we repeated the experiment where the location is given by a segmentation.

For the distinction between object and surround using boxes, results are almost identical to the ones presented in this paper. For the normal situation this should come as no surprise given the similarities in the visualisation of the pixel-wise contribution in Fig. 3. For the ideal situation recall that a box already includes some surround. Descriptors within the box will measure a bit more of the surround but not significantly. Descriptors outside the box can measure slightly bigger parts of the object but most of the descriptor is still used for measuring the surround. Overall, using 4×4 SIFT yields figures nearly identical to the ones presented in Figs. 6 and 7 and does not affect our conclusions.

For the distinction between interior and border, we carefully made our interior such that it does not contain any border. For the larger descriptor, this means that we had to make our interior smaller such that is was only on average 35% of the total size of the object. In both the normal and ideal situation all tendencies are very similar: The border alone becomes slightly more predictive of the class, while the predictive power of the interior remains approximately the same for all classes except *bike* and *plant*. For *bike* this is because half of the classes lose all of its interior. Again, for the animal classes *cat, cow, dog*, and *sheep* adding the border to the interior does not yield significant improvements over using the interior alone.

Summarised, results are almost the same when using 4×4 SIFT. Hence our conclusions remain valid for this larger, more commonly used descriptor.

5 Discussion and Conclusion

This paper investigated the visual extent of an object in terms of the object and its surround, and in terms of the object interior and the object border. Our investigation was performed from two perspectives: The normal situation where the location of the objects are unknown, and an ideal situation with known object locations.

For the normal perspective we visualised in Sect. 4.1.1 how the Bag-of-Words framework classifies images. These visualisations indicate that the support for the classifiers is found throughout the whole image occurring indiscriminately in both the object and its surround, supporting the notion that context facilitates image classification (Divvala et al. 2009; Oliva and Torralba 2007). While for some classes with a highly varying surround Bag-of-Words learns to ignore the context, as observed by Zhang et al. (2007), this does not generalise to all classes. We found that the role of the surroundings is significant for many classes, to the point where for *boat* and *bottle* they are even more important for recognition than the object itself. For *boat* the object area is even a negative indicator of its presence.

At the same time, we have demonstrated in Fig. 7(b) that when the object locations are known a priori, the surroundings do not help to increase the classification performance significantly. After ideal localisation, regardless of the size of the object, the object appearance alone predicts its presence equally well as the combination of the object appearance and the surround.

We showed that no visual words uniquely describe only object or only surround. However, by making the distinction between object and surround explicit using the object locations, performance increases significantly by 0.20 MAP. This suggests that modelling the object location can lead to further improvements within the Bag-of-Words framework, where we see the work of Harzallah et al. (2009) as a promising start.

Regarding the surround the following view arises. The surroundings are indispensable to distinguish between groups of classes: furniture, animals, and land-vehicles. When distinguishing among the classes within one group the surroundings are a source of confusion.

Regarding the object features, we have observed differences how classes are being recognised: (1) For the physically rigid *aeroplane, bicycle, bus, car*, and *train* classes interior and exterior boundaries are important, while texture is not. (2) The non-rigid animals *dog, cat, cow,* and *sheep* are recognised primarily by their fur while their projected shape varies greatly. While SIFT feature values respond to interior boundaries, exterior boundaries, and texture at the same time, the recognition differences suggest that using more specialised features could be beneficial.

Bag-of-Words with SIFT measure texture, interior object boundary fragments, and shape boundary fragments as local details. For identifying the context of an image this is adequate, especially considering that context partially consists of shapeless mass-goods such as grass, sky, or water. In contrast, for objects features more spatial consistency could help. This suggests that future features would render more improvements on recognising objects than on recognising context. Intuitively, this means that when the exact object location is known, context helps less for recognition than our experiment in Fig. 7(b). This is consistent with the observation by Biederman (1981) in human vision that objects viewed in isolation are recognised as easily as objects in proper context.

References

Agarwal, S., Awan, A., & Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(11), 1475–1490.

Bar, M. (2004). Visual objects in context. *Nature Reviews. Neuroscience*, *5*, 617–629.

Biederman, I. (1981). On the semantics of a glance at a scene. In *Perceptual organization* (pp. 213–263). Hillsdale: Lawrence Erlbaum.

Bishop, C. M. (2006). *Pattern recognition and machine intelligence*. Berlin: Springer.

Blaschko, M. B., & Lampert, C. H. (2009). Object localization with global and local context kernels. In *British machine vision conference*.

Burl, M. C., Weber, M., & Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. In *European conference on computer vision*.

Carbonetto, P., de Freitas, N., & Barnard, K. (2004). A statistical model for general contextual object recognition. In *European conference on computer vision*. Berlin: Springer.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *ECCV international workshop on statistical learning in computer vision*, Prague.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE conference on computer vision and pattern recognition*.

Divvala, S. K., Hoiem, D., Hays, J. H., Efros, A. A., & Herbert, M. (2009). An empirical study of context in object detection. In *IEEE conference on computer vision and pattern recognition*.

Everingham, M., van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*, 303–338.

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *IEEE conference on computer vision and pattern recognition*.

Fulkerson, B., Vedaldi, A., & Soatto, S. (2009). Class segmentation and object localization with superpixel neighborhoods. In *IEEE international conference on computer vision*.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42.

Gould, S., Fulton, R., & Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. In *IEEE international conference on computer vision*.

Harzallah, H., Jurie, F., & Schmid, C. (2009). Combining efficient object localization and image classification. In *IEEE international conference on computer vision*.

Hoiem, D., Efros, A. A., & Hebert, M. (2008). Putting objects in perspective. *International Journal of Computer Vision*, *80*, 3–15.

Jiang, Y. G., Ngo, C. W., & Yang, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM international conference on image and video retrieval* (pp. 494–501). New York: ACM Press.

Jurie, F., & Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *IEEE international conference on computer vision*.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE conference on computer vision and pattern recognition*, New York.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110.

Maji, S., Berg, A. C., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *IEEE conference on computer vision and pattern recognition*.

Malisiewicz, T., & Efros, A. A. (2007). Improving spatial support for objects via multiple segmentations. In *British machine vision conference*, September 2007.

Malisiewicz, T., & Efros, A. A. (2009). Beyond categories: the visual memex model for reasoning about object relationships. In *Neural information processing systems*.

Marszałek, M., Schmid, C., Harzallah, H., & van de Weijer, J. (2007). Learning representations for visual object class recognition. In *ICCV Pascal VOC 2007 challenge workshop*.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1615–1630.

Moosmann, F., Triggs, B., & Jurie, F. (2006). Fast discriminative visual codebooks using randomized clustering forests. In *Neural information processing systems* (pp. 985–992).

Nedović, V., & Smeulders, A. W. M. (2010). Stages as models of scene geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 1673–1687.

Nowak, E., Jurie, F., & Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *European conference on computer vision*.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*, 520–527.

Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., & Belongie, S. (2007). Objects in context. In *International conference on computer vision* (pp. 1–8).

Shotton, J., Winn, J., Rother, C., & Criminisi, A. (2009). Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, *81*, 2–23.

Singhal, A., Luo, J., & Zhu, W. (2003). Probabilistic spatial context models for scene content understanding. In *IEEE conference on computer vision and pattern recognition*.

Sivic, J., & Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. In *IEEE international conference on computer vision*.

Smeaton, A. F., Over, P. & Kraaij, W. (2006). Evaluation campaigns and TRECVID. In *ACM SIGMM international workshop on multimedia information Retrieval*.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(12), 1349–1380.

Tahir, M. A., van de Sande, K., Uijlings, J., Yan, F., Li, X., Mikolajczyk, K., Kittler, J., Gevers, T., & Smeulders, A. (2008). UVA and surrey @ Pascal VOC 2008. In *ECCV Pascal VOC 2008 challenge workshop.*

Tuytelaars, T., & Schmid, C. (2007). Vector quantizing feature space with a regular lattice. In *IEEE international conference on computer vision*.

Uijlings, J. R. R., Smeulders, A. W. M., & Scha, R. J. H. (2009). What is the spatial extent of an object? In *IEEE conference on computer vision and pattern recognition*.

Uijlings, J. R. R., Smeulders, A. W. M., & Scha, R. J. H. (2010, in press). Real-time visual concept classification. *IEEE Transactions on Multimedia*. http://dx.doi.org/10.1109/TMM.2010.2052027

Ullah, M. M., Parizi, S. N., & Laptev, I. (2010). Improving bag-of-features action recognition with non-local cues. In *British machine vision conference*.

van de Sande, K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 1582–1596.

Vedaldi, A., & Zisserman, A. (2010). Efficient additive kernels via explicit feature maps. In *IEEE conference on computer vision and pattern recognition*.

Wolf, L., & Bileschi, S. (2006). A critical view of context. *International Journal of Computer Vision*, *69*, 251–261.

Zhang, J., Marszałek, M., Lazebnik, S., & Schmid, C. (2007). Local features and Kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, *73*(2), 213–238.