International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

The Visual Extent of an Object

Suppose we have the bounding boxes around objects

J.R.R. Uijlings[1], **A.W.M. Smeulders**[1], **R.J.H. Scha**[2]

[1] Institute for Informatics, ISIS Lab,
Science Park 107, 1098 XG, Amsterdam, The Netherlands
e-mail: JRR.Uijlings@uva.nl
[2] Institute for Logic, Language and Computation,

Received: date / Accepted: date

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. This paper investigates the visual extent of an object from two angles: (a) Not knowing the object location, we determine where in the image the support for object classification resides. (b) Assuming an ideal box around the object we evaluate the relative contribution of the object interior, object border, and surround. In (a) we find that the surroundings contribute significantly to object classification where for *boat* the object area contributes negatively. In (b) we find that the surroundings no longer contribute. Comparing (a) and (b) we find that with ideal object localisation there is a considerable gain in classification accuracy to be made.

Key words Content Based Image Retrieval Visual Extent Context

1 Introduction

It is widely acknowledged that the visual extent of an object extends beyond the object itself (*e.g.* [2,3,28,40]). Nevertheless, in the early days of computer vision it was thought that the visual extent of an object is precisely confined to the object silhouette. 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 [34]. 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 localization of the object was to use local region descriptors in a specific spatial arrangement [1,5,11]. 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 [32], 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*. [7] generalized 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 [18, 26]. Therefore these words are no longer extracted at salient points but on a dense, regular grid. This has caused last notion of object location to be lost in the Bag-of-Words representation. This is the state-of-the-art of image classification in 2009 [10,33].

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 extends our CVPR paper [37]. Specifically, we investigate what is the influence of the surroundings, what is the influence of the object borders, and what is influence of the object interior for object classification?

2 Related Work

The influence of context on recognition was researched earlier in human vision. Most notably, Biederman [3] 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 [28] 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 [40], 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*. [9] identifies several types of context as used in the computer vision community. These include Local Pixel Context [6,8, 12,14,30], 2D scene gist context [27], 3D geometric context [16,25], and semantic context [29,31]. 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 (*e.g*. [6] or Conditional Random Fields [6,30] and we do not use a scene label, but we directly use the features which we extract from the image.

Zhang *et al*. [41] 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 2007 dataset, we will challenge this finding and investigate whether the influence of context in the Bag-of-Words framework is significant.

Tuytelaars and Schmid [36] 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.

Harzallah *et al*. [15] 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. In this paper we provide an upper bound of retrieval performance when the object is localised, showing that using isolated object features is not only good idea but also that there is plenty of room for further improvement in this direction.

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 [33] and the Pascal VOC Classification task for retrieving images [10]. 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 Figure 1. The first line is the *normal* situation where we apply a visual concept detection algorithm and determine which image parts contribute how much to the identification of the target object. The second line is the *ideal* situation in which 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 one 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 object locations in the form of ground truth bounding boxes to isolate the object from its surround in both the training and test set in both lines of our investigation. In the normal line we do this after classification and in the ideal line we do this beforehand. Note that in both cases we only use the *location* of the bounding boxes and not their labels. This means that if there is more than one object in the image we leave it to the classifier to select the correct bounding box. This reflects the situation where one is able to successfully segment the objects within an image. Using this strategy instead of always selecting the correct bounding box does not influence the general observations made in this paper (data not shown).

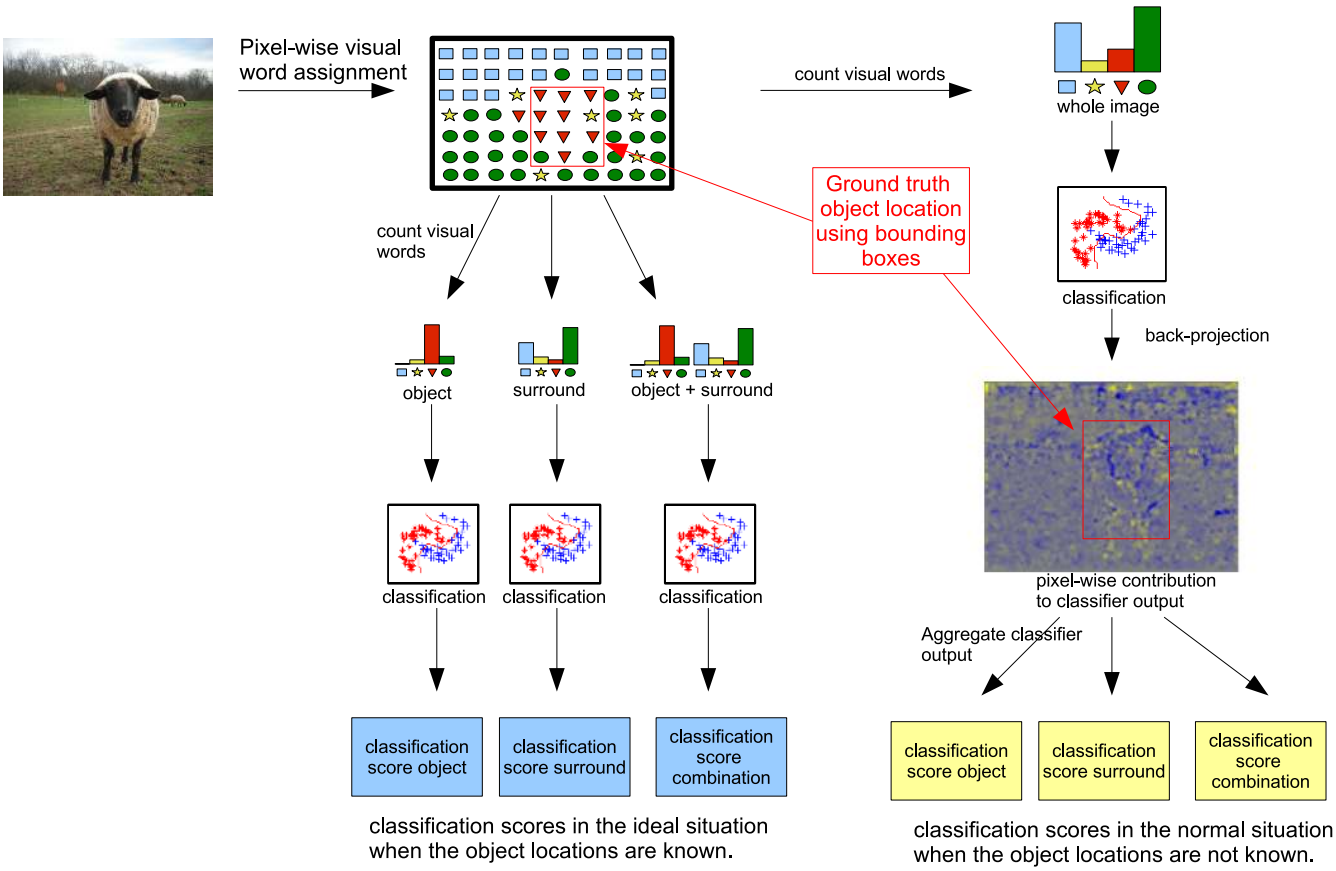


Fig. 1 The two main lines of our analysis: The normal line on the right uses the ground truth object locations to divide the image into object and surround before classification. The ideal line (on the left) first classifies the image, projects the classification score back on the image and then aggregates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

As we do not have the ground truth to distinguish between object interior and object border, we need to determine this ourselves. We consider the boundary between the object interior and the object border in terms of the object interior box which splits the complete object into the two desired parts. We define the 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 is complementary information. We can now find the optimal interior box by representing the object through its border and interior and find the box which maximises classification accuracy.

3.1 Dataset

We choose the large-scale, well-studied Pascal VOC 2007 dataset for our analysis. This comes at the expense of the quality of the annotations: we use a bounding box around the object rather than segmented objects. By this choice, we emphasize quantity of annotations over the quality of annotations. Furthermore, the widespread use of this dataset allows for a better interpretation and comparison with respect to other work.

The Pascal VOC 2007 challenge consists of 9963 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 dataset is split into two predefined train and test sets of size 5011 and 4952 images respectively.

The classification performance of the Pascal VOC dataset is measured by the Average Precision of a ranked list, 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 $X = \{x_1, \cdots, x_n\}$. Finally, f_c is a function which returns the 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

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

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. The use of the CAMP in our experiments helps in determining the cause of accuracy loss or gain.

3.3 Bag-of-Words Framework

A condense overview of our Bag-of-Words implementation [38] is given in Table 1. We sample small regions at each pixel which is an extreme form of sampling using a regular, dense grid. [18,26]. From these regions we extract SIFT [20] and four colour SIFT variants [39] which have been shown to be superior for image retrieval [23,39,41]. Thus we use intensity-based SIFT, opponent-SIFT, rg-SIFT (normalized 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 section 4.1. The size of each SIFT patch is 8 by 8 pixels.

For the creation of a visual vocabulary we use a Random Forest [24] 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 [24,38]. 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 randomized trees algorithm [13], using 500,000 labelled descriptors sampled randomly from the training set, where the labels are obtained from the annotation at image level.

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

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 bounding boxes denoting the object locations.

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 [17,41]. 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 section 3.4.1. By taking the square root of the visual word histograms we compensate for high frequent visual words, which makes the Histogram Intersection kernel almost as good as the χ^2 kernel [38]. In fact, by sampling visual words every pixel we found no difference in performance between the χ^2 kernel and the histogram intersection kernel.

The original Bag of Words framework is orderless. Therefore Lazebnik *et al.* [19] introduced a weak spatial order by using their spatial pyramid, in which an image is divided into regular subregions. Codebook 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 pyramid division on this dataset by several researchers [22, 35,38]. 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.

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 bounding boxes to determine how much each image part (*i.e.* surround, object, object interior, object border) contributes to the classification. When an image contains instances of multiple object classes, we use the one object class with the highest classifier output to make the distinction into object, surround, object interior, and object border. This allows us to create a partitioning for both target and non-target images, which in turn enables us to calculate the Average Precision over the whole dataset (instead of over only 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 and is done as follows.

The classification function for a Support Vector Machine can be written as [4]

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

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 [21]

$$h(\mathbf{x}) = b + \sum_{j=1}^{m} \alpha_j t_j k(\mathbf{x}, \mathbf{z}_j) \tag{5}$$

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

As the outer sum in equation 6 is over the visual words, the contribution per visual word channel w_i is calculated as

$$w_i = \sum_{j=1}^{m} \alpha_j t_j \min(x_i, z_{ij}). \tag{7}$$

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 Figure 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 bounding boxes 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 contains multiple object classes we let the classifier decide which class is used to divide the image into object, surround, object interior, and object border. The class with the highest object score is considered the target object class and is used to divide the image. This strategy allows us to create visual word representations for both target and non-target images, enabling calculating the Average Precision over the whole dataset.

4 Results

4.1 Classification without knowing the object location

For our normal Bag-of-Words system where we do 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 [15] and [39], which are respectively 0.60 MAP and 0.61 MAP. To enable back-projection with equation 7 we use the Histogram Intersection kernel instead of the widely accepted χ^2 kernel [15,17,39,41]. 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 [15,39] may 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, our experiments demand the use of small SIFT descriptors to minimize the overlap between object and surround descriptors.

The confusion matrix of the normal Bag-of-Words system is shown in Figure 2. 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 Figure 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.

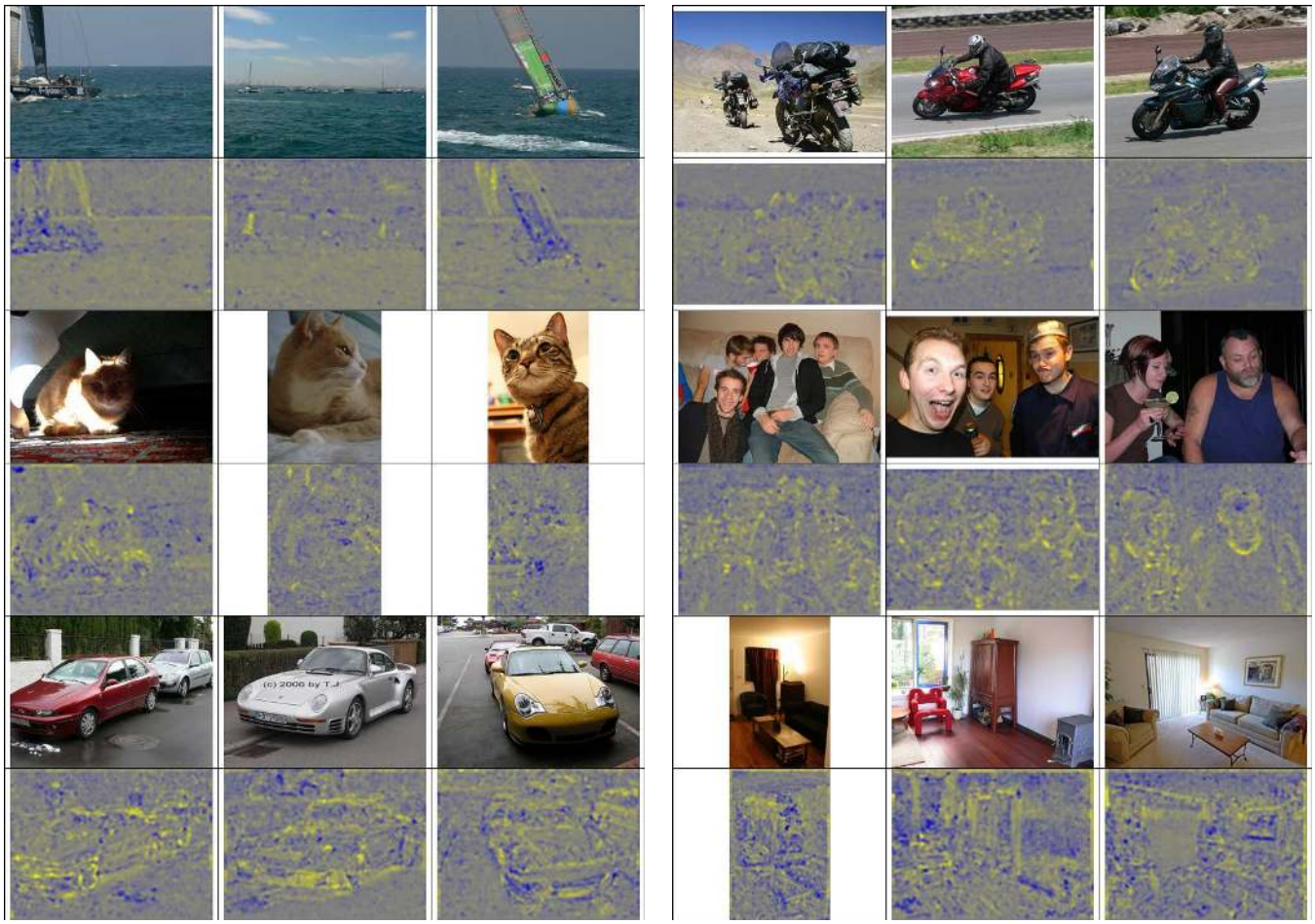


Fig. 3 Pixel-wise contribution to the classification for top ranked images for the categories *boat, cat, car, motorbike, person*, and *sofa*. Dark-blue means a negative and light-yellow means a positive contribution to the classifier. Notice that high positive or high negative contributions are often located on small details.

4.1.1 Localising the Classifier Contributions We now investigate qualitatively where the Bag-of-Words classifier obtains the evidence to classify images. We use the method described in section 3.4.1 and show results for top-ranked images of the classes *aeroplane, boat, cat, car, person*, and *sofa* in Figure 3.

First we observe that 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 water-sky transition 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 as a hole in the water and is 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 not ideal. Furthermore the shape of the cat is not important. We see similar behaviour for the other animal classes except *horse*, whose shape of legs are an important cue (data not shown). This suggests that most animals are recognized 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, and window-hood boundary. The impor-

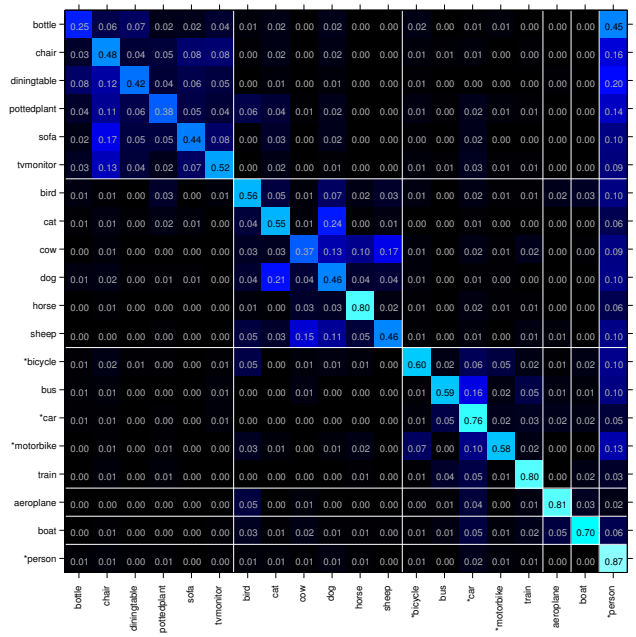


Fig. 2 Average Precision Confusion Matrix (CAMP) of the normal situation where the object locations are unknown. Notice the within-category confusion in the furniture, animal, and land-vehicle classes.

tance 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 confusion with the *bicycle* class (data not shown).

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 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 exterior of the object. We compare this with the results of normal situation where the object location is not known.

First we need to determine the location of the object interior boxes to make the distinction between the interior and border as described in section 3. We determine these locations in the ideal situation on half of the Pascal dataset (`train + val`). The interior box is defined as being a certain factor smaller than the object box while its centre pixel remains the same. We shrink the object box to 10% to 90% of its original size with increments of 10%. We create an object representation by concatenating the visual word histograms of the interior and border. The optimal classification accuracy in terms of Mean Average Precision is achieved by shrinking the interior box in the range of 50%-70% of the size of the complete object box. Within this range all results presented in this paper are similar (data not shown). We show results for shrinking the interior box to 60% of the size of the complete object box, which means the object interior covers 36% of the *surface area* of the complete object while the object border covers 64%.

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 0.57 MAP, the accuracy of the ideal situation when making the distinction between object and surround is 0.73 MAP. When creating separate representations for the surround, object interior, and object border performance increases to 0.77 MAP. This shows that the potential gain of knowing the object locations but not their labels is 0.20 MAP in this dataset.

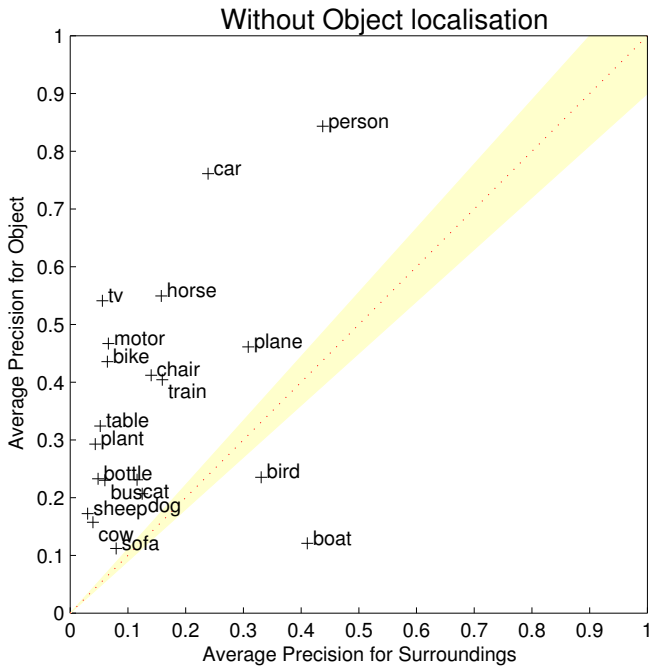
The huge difference between the accuracy without and without knowing the object location shows that the classifier cannot distinguish if visual words belong to the object or surround. We investigate the cause by determining for each visual word the probability that it occurs in any object and that it occurs in the surround, which is visualised in Figure 7. This graph shows that only a few words have a larger than 90% probability of describing background. These words describe homogeneous texture (data not shown). In contrast, there is only a single word that describes in more than 66% of the cases an object and that word occurs only 21 times in the whole dataset. This means that no visual words exclusively describes objects and that these visual words are less specific than is generally thought.

4.3 Discussion on Object versus Surround

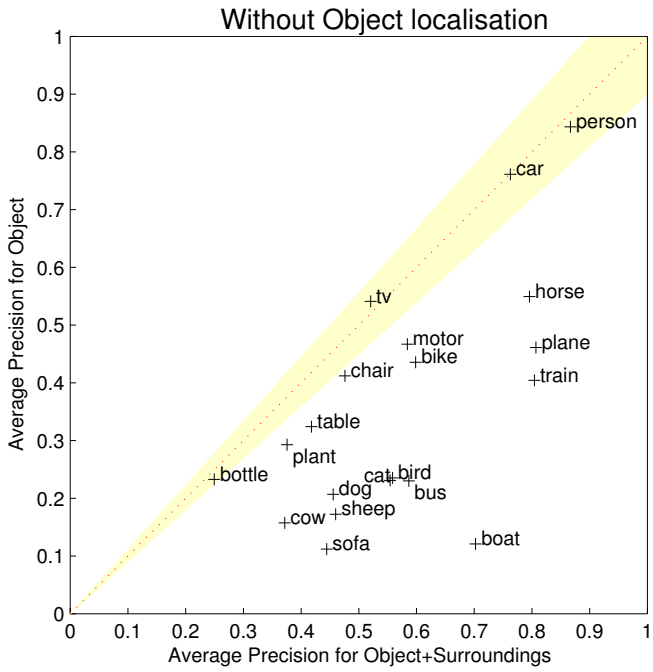
We now proceed to discuss the relative influence of the object and its surroundings. Figure 5 plots the Average Precision for the object against the surround and against the combination of the object and surround for the normal situation where the object location is unknown, Figure 6 plots the same for the ideal setting where the object location is known.

In Figure 5(a) one can see that for *boat* and *bird* the surroundings are more used than the object for classification in the normal situation. For *boat* this confirms that it is recognised as a hole in the water as seen in Figure 3.

The retrieval performance when using only the surround is very low for most classes in the normal setting, except for *boat, bird, person*, and *plane*. In contrast, when training and learning on the isolated surroundings, Figure 6(a) shows that most classes can be retrieved reasonably well. The difference is especially large for *train* and *horse*. Thus, while the sur-

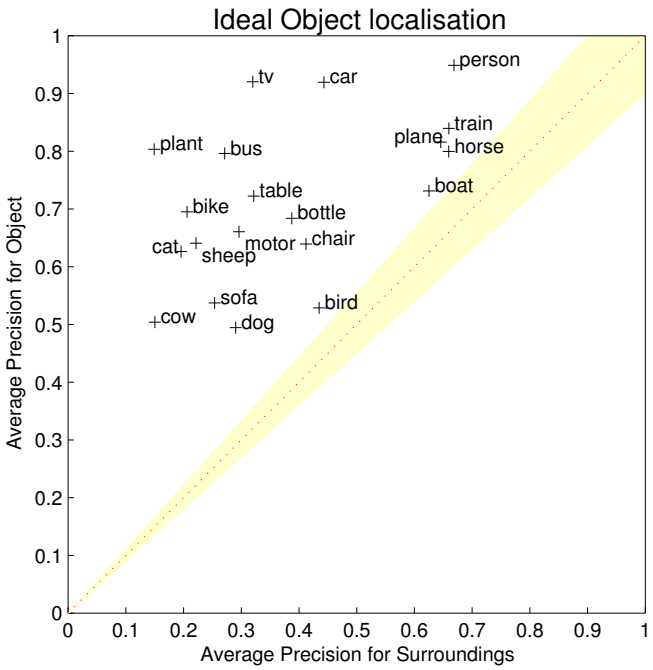


(a)

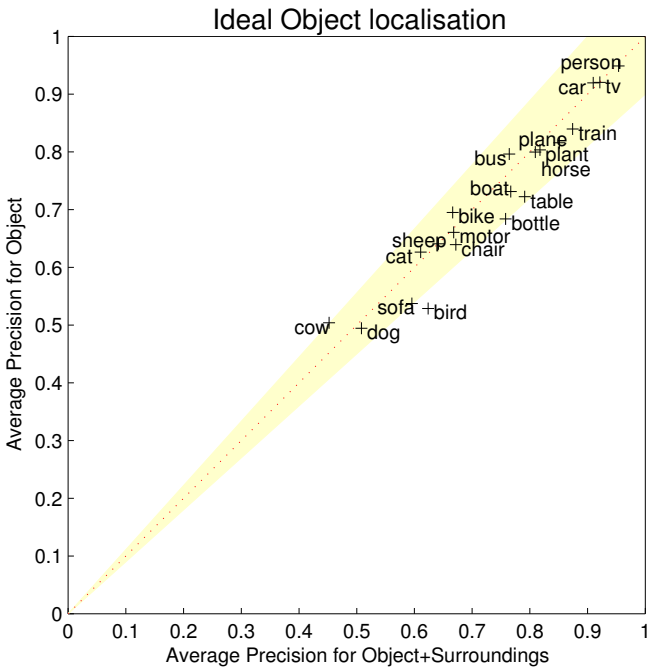


(b)

Fig. 5 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 *bird* and *boat* the surround is more important than the object itself. (b) The object versus the combination of object and surround. For *bottle, tv/monitor, car*, and *person* the performance of the object is very similar to the combination, suggesting that these classes are context-free.



(a)



(b)

Fig. 6 The retrieval performance of the object, surround, and its combination in the ideal setting where the object locations are considered known. (a) the surround versus the object. For all classes the object is more important than the surround. For all classes performance increases significantly over Figure 5(a) (b) The object versus the combination of object and surround. For most classes performance is similar for the object and the combination. This means that if the object location is considered known, the surround adds little information.

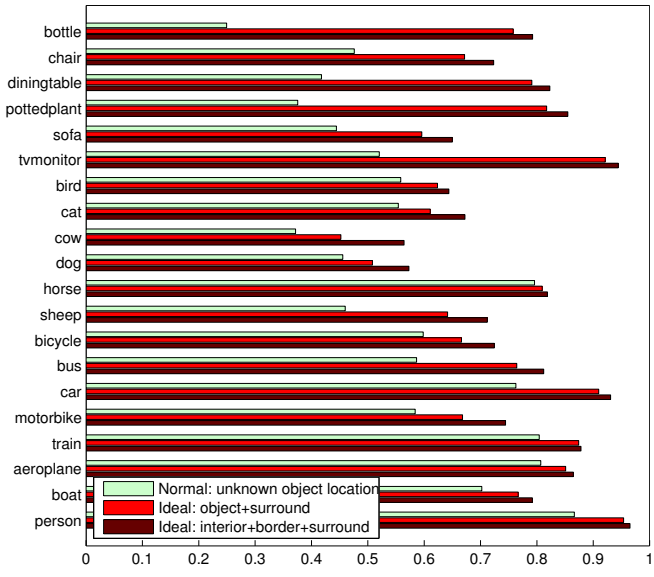


Fig. 4 A comparison of the normal situation when the object location is unknown and the ideal situation where the object location is known. Accuracy over all classes for the normal situation is 0.57 MAP, for object+surround this is 0.73 MAP, and for interior+border+surround this is 0.77 MAP.

roundings contain information, it is not the focus of the classifier.

In Figure 5(b) we see that the combination of object and surround is much better than using the object alone for most classes. This is not surprising as the classifier was learned on the combination. However, for the classes *bottle, tv/monitor, car*, and *person* the performance of the combination is equal to using only the object. This means that the classifier learns to ignore the surroundings, suggesting that these classes are context-free.

When the objects are considered localised in Figure 6(b), for all classes except *bird* and *bottle*, using surroundings in addition to the object does not yield much improvement over using the object alone. Interestingly, this agrees with the research on human vision of Biederman [3], who found that objects viewed in isolation are recognised equally well as objects viewed in proper context.

In the idealized setting when the object is considered localized, we also analyse the confusion matrices of using only object and context descriptors in Figure 8. The confusion matrix of using only surround in Figure 8(a) looks similar to the confusion matrix of the normal setting in Figure 4.1. 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 a race circuit and the bus-context is dominated by urban environments, whereas the *car* occurs in both.

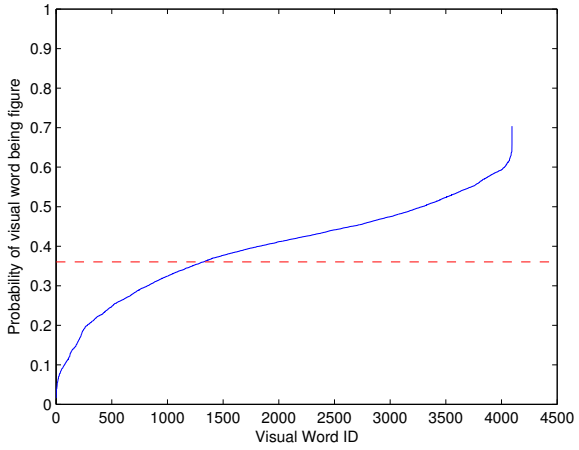


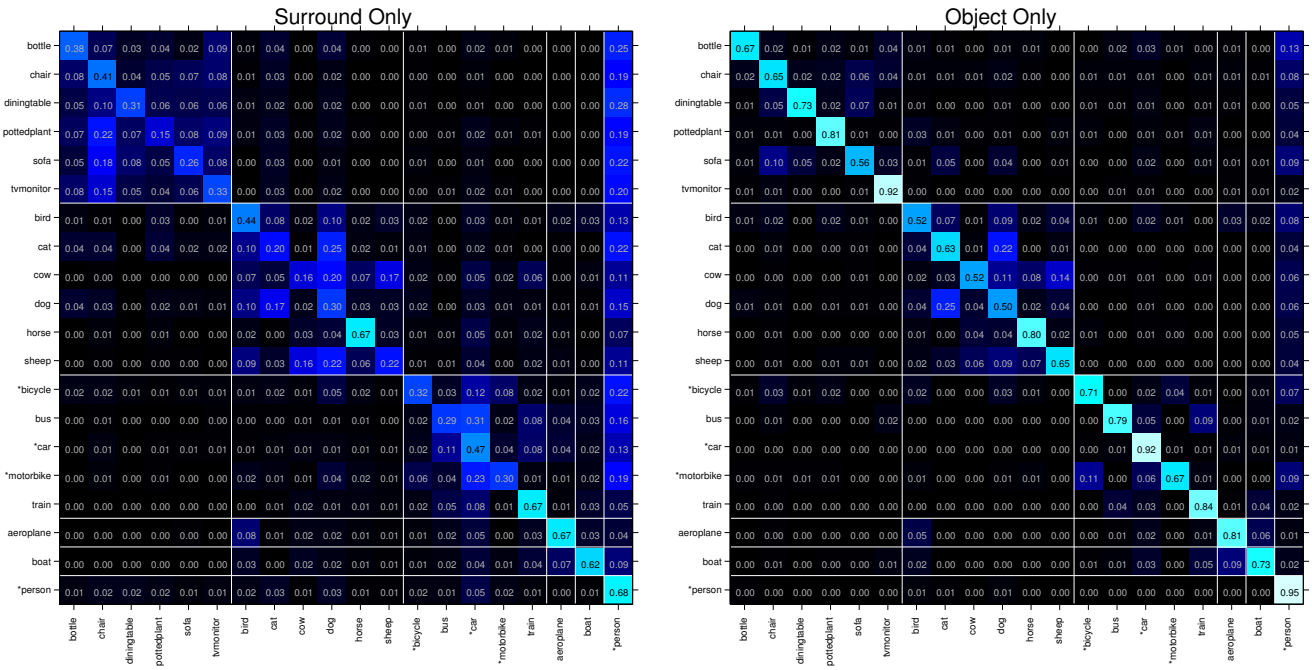
Fig. 7 The probability of each visual word belonging to an object. The dotted red line is the prior probability. Contrary to general belief, visual words are not very object specific as no visual word occurs more than 70% on an object.

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. The only exception is that without the surround bicycles are far more confused as motorbikes, but not vice versa. 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 the next section we will see what causes most confusion: fur or shape.

To conclude, in the normal situation where the object location is unknown the surroundings contribute significantly to classification for all classes but *bottle, car, person*, and *tv/monitor*, which are context-free. 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.* [41] that the surroundings are not important for Bag-of-Words still holds for *car* and *person*, but do not generalise to other classes in larger datasets. In contrast, when the object locations are known, the surroundings add little additional information which is in accordance with human vision [3]. 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 Interior versus Border

Now we discuss the relative influence of the interior and the border of the object. Figure 9 plots the Average Precision for the interior against the border and against the combination of the interior and border for the normal situation where the object locations are unknown, Figure 10 plots the same for the ideal situation with known object location.



(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. What is noticeable is that when 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.

For all classes the influence of the object interior and boundary show similar tendencies in both the normal and ideal setting. For the classes *boat, sofa*, and *tv/monitor* the border is yields better results than the interior in Figure 9(a) and 10(a). For *tv/monitor* this is because its interior can take any appearance and Figure 3 shows that for *sofa* the classifier contributions lie mainly on vertical and diagonal lines. However, for *boat* the preference for the border is most likely because it often contains water.

In contrast, for the classes *dog, cat, cow, sheep*, and *motor* the object interior is more important than the object borders. For the animals this makes sense as they are non-rigid objects which can be found in a variety of poses, hence their shape information is volatile and unreliable while their fur is a stable source of evidence. For *cat* this was observed earlier in Figure 3. For *motor* the behaviour is more surprising. However, inspection of the confusion matrices show that the *motorbike* boundaries perform worse because they get confused with *bicycle* boundaries, which in turn is caused by the wheel shape as seen in Figure 3. Note that this confusion works only one way: for *bike* the pixel-wise classifier contributions show that the classifier focuses on the spokes (data not shown).

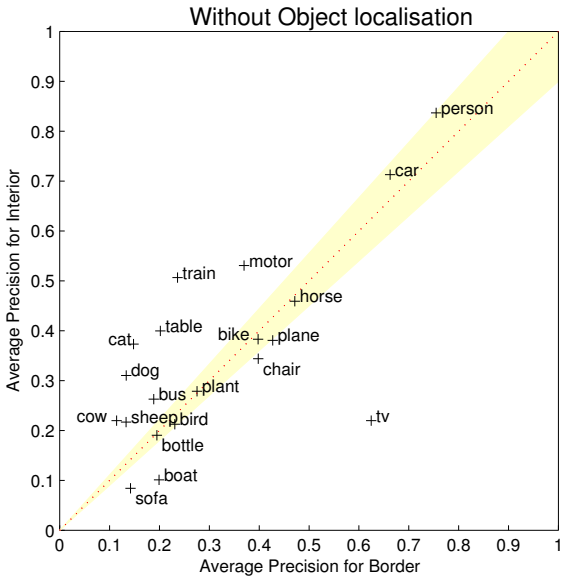
For the physically rigid *bus, car*, and *aeroplane* classes, both the interior and border are about equally important. The pixel-wise classifier contributions show that mainly boundary edges are important as can be seen for *car* in Figure 3. These edges include both shape boundaries and interior boundaries.

Finally, in Figure 9(b) one can see that the classes *dog, cat, cow, sheep*, and *motor* for which the object interior is more important than the object border, the object interior is also more important than the combination of the two. In contrast, in Figure 10(b) where the object locations are known, classification is slightly better when using both the interior and border for the classes *cow, sheep*, and *motor*. This means that the border for these classes does contain discriminative information. Observations are similar for the classes *boat, sofa*, and *tv/monitor* whose border is more important (data not shown).

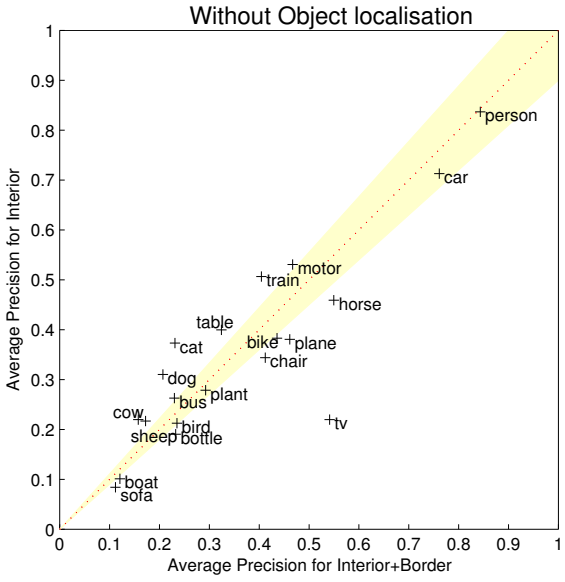
5 Conclusions

This paper investigated the visual extent of an object in terms of its surround, its interior and its border 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 situation we visualised how the Bag-of-Words framework classifies images. As expected, 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 [9,28]. The role of the surroundings is significant for most classes to the point where,

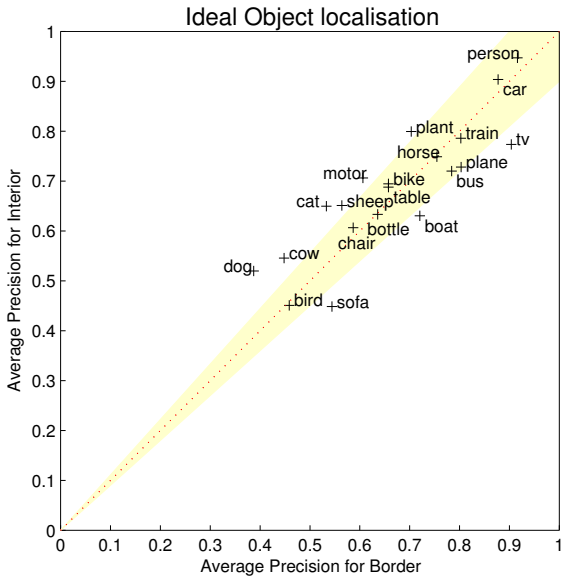


(a)

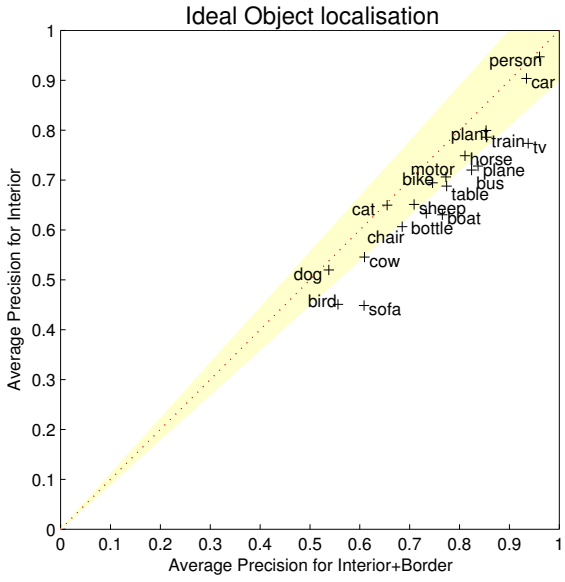


(b)

Fig. 9 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. *dog, cat, cow, train, table, sheep, motor*, and *bus* are better recognised by their interior, *tv/monitor, boat*, and *sofa* are better recognised by their border. (b) Object interior versus the combination. For most classes that are better recognised by their interior using also the border information lowers performance.



(a)



(b)

Fig. 10 The retrieval performance of the interior, border, and their combination in the idealised setting where the object locations are considered known. (a) As in Figure 9(a), the classes *dog, cat, cow*, and *motor* are best recognised by their interior, *tv/monitor, boat*, and *sofa* are better recognised by their border. (b) object interior versus the combination. The combination of interior and border is always as good or better than using the interior alone.

surprisingly, for *boat* and *bird* they are even more important for recognition than the object itself. For *boat* the object area is even a negative indicator of its presence. Thus, the observation of Zhang *et al*. [41] that surroundings are unimportant in the Bag-of-Words framework does not generalise. For the classes *bottle, car, person*, and *tv/monitor* the framework ignores the surroundings, suggesting that these classes are context-free.

In contrast, in the ideal case where the object bounding boxes are known a priori, using the surroundings in addition to the object does *not* help to increase classification performance significantly. The object is classified purely by its own appearance (and a marginal surround inside the bounding box). This is quite different from the normal situation where the support for the classification is scattered over the image as argued above. Yet it is consistent with the observation by Biederman [3] in human vision that objects viewed in isolation are recognised as easily as objects in proper context.

In general, the surroundings help to distinguish between *groups* of classes: *furniture*, *animals*, and *land-vehicles* all have distinct surroundings. When distinguishing between the classes of 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, bus*, and *car* classes interior and exterior boundaries are important, while texture is not. (2) For the classes *sofa* and *tv/monitor* only the exterior boundary is important as their interior can take any appearance. (3) The non-rigid animals *dog, cat, cow,* and *sheep* are recognised primarily by their fur while their projected shape varies highly. 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 is beneficial.

When the object locations are unknown our retrieval performance is 0.57 MAP, whereas when object locations are known retrieval performance increases to 0.77 MAP, which is a significant difference of 0.20 MAP. This difference means that without object locations the classifier is unable to learn the origin of visual words. Further examination shows that no visual words are specific for objects. Instead one could model the object location: Harzallah *et al*. [15] fused the output of their object *localisation* system with their object *classification* and improved classification accuracy by 0.035 MAP to 0.635 MAP. Being a positive result in its own right, on the basis of results in this paper we conclude that there is still room for improvement in using object locations to improve image classification.

References

1. S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490, 2004.
2. M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5:617–629, 2004.
3. I. Biederman. *Perceptual Organization*, chapter On the semantics of a glance at a scene, pages 213–263. Lawrence Erlbaum, Hillsdale, New Jersey, 1981.
4. C.M. Bishop. *Pattern Recognition and Machine Intelligence*. Springer Science+Business Media, LLC, 2006.
5. M.C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *European Conference on Computer Vision*, 1998.
6. P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proceedings of the Nineteenth Annual SAS Users Group International Conference*. Springer, 2004.
7. G. Csurka, C.R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, Prague, 2004.
8. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005.
9. S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros, and M. Herbert. An empirical study of context in object detection. In *IEEE conference on Computer Vision and Pattern Recognition*, 2009.
10. M. Everingham, L. van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Jounal of Computer Vision*, In press.
11. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
12. B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *IEEE International Conference on Computer Vision*, 2009.
13. P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.
14. S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *IEEE International Conference on Computer Vision*, 2009.
15. H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *International Conference on Computer Vision*, 2009.
16. D. Hoiem, A.A. Efros, and M. Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 2007.
17. Y.G. Jiang, C.W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM International Conference on Image and Video Retrieval*, pages 494–501. ACM Press New York, NY, USA, 2007.
18. F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *IEEE International Conference on Computer Vision*, 2005.
19. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, New York, 2006.
20. D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.
21. S. Maji, A.C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
22. M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning representations for visual object class recognition. ICCV Pascal VOC 2007 challenge workshop., 2007.

23. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.
24. F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Neural Information Processing Systems*, pages 985–992, 2006.
25. V. Nedović and A.W.M. Smeulders. Stages as models of scene geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.
26. E. Nowak, F. Jurie, and B. Triggs. Sampling Strategies for Bag-of-Features Image Classification. In *European Conference on Computer Vision (ECCV)*, volume 3954, page 490. Springer, 2006.
27. A. Oliva and A Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
28. A. Oliva and A Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11:520–527, 2007.
29. A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.
30. J. Shotton, J. Winn, C. Rother, and A. Criminisi. 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, 2009.
31. A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 1:235–241, 2003.
32. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer VIsion*, 2003.
33. A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*, 2006.
34. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
35. M.A. Tahir, K. van de Sande, J. Uijlings, F. Yan, X. Li, K. Mikolajczyk, J. Kittler, T. Gevers, and A. Smeulders. Uva and surrey @ pascal voc 2008. ECCV Pascal VOC 2008 challenge workshop., 2008.
36. T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *IEEE International Conference on Computer Vision*, 2007.
37. J.R.R. Uijlings, A.W.M. Smeulders, and R.J.H. Scha. What is the spatial extent of an object? In *CVPR*, 2009.
38. J.R.R. Uijlings, A.W.M. Smeulders, and R.J.H. Scha. Real-time visual concept classification. *IEEE Transactions on Multimedia*, In press, 2010.
39. K.E.A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. (in press).
40. L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 69:251–261, 2006.
41. J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision*, 73(2):213–238, 2007.