Spatial Weighting for Bag-of-Features

Marcin Marszałek   Cordelia Schmid

INRIA Rhône-Alpes, LEAR - GRAVIR
665 av de l'Europe, 38330 Montbonnot, France
Marcin.Marszalek@inrialpes.fr   Cordelia.Schmid@inrialpes.fr

Abstract

This paper presents an extension to category classification with bag-of-features, which represents an image as an orderless distribution of features. We propose a method to exploit spatial relations between features by utilizing object boundaries provided during supervised training. We boost the weights of features that agree on the position and shape of the object and suppress the weights of background features, hence the name of our method— "spatial weighting". The proposed representation is thus richer and more robust to background clutter. Experimental results show that our approach improves the results of one of the best current image classification techniques. Furthermore, we propose to apply the spatial model to object localization. Initial results are promising.

1. Introduction

The recognition of object categories is one of the most challenging problems in computer vision, especially in the presence of pose changes, intra-class variation, occlusion and background clutter. Methods based on sparse local features [1, 3, 5] and bag-of-features [23, 24] were proposed to deal with pose changes and intra-class variations. They have shown to give excellent results. However, they are sensitive to background clutter, because they cannot distinguish between objects and background. Efforts have been made to overcome this problem by using feature selection [3], boosting [19] or designing novel kernels with high discriminative power [9, 16]. Robustness to occlusions was improved by introducing similarity measures based on partial matching (EMD distance [21]) or histogram comparison (χ^2 distance [10]). On those distances robust Gaussian kernels for Support Vector Machines (SVM) [22] were built. However, there still seems to be a strong potential for improving background clutter robustness of bag-of-keypoints representation [25].

It has been shown that considering spatial relationships between features, which are ignored by the standard bag-of-keypoints representation, may lead to high recognition results [12]. This motivates us to extend the original bag-of-keypoints representation to incorporate spatial information. It was also shown that interest points can generate accurate hypotheses about localization of the object in the image [13, 15]. Extending those ideas, we introduce a method in which the features that agree on the localization and shape of the object boost the importance of each other. We name our technique "spatial weighting". We test the performance of our approach on the PASCAL Visual Object Classes Challenge data set [4] (see fig. 1). We evaluate a state-of-the-art method of this challenge [4, 25] and show that applying the proposed method can improve results.

Our approach uses all the information provided during supervised training, i.e., not only the image label, but also the object localization. Traditional bag-of-features approaches can use during training either all-image information (weakly supervised setting) or object-only information. As was shown in [25], both methods have drawbacks. We will overcome this by employing all the information given during the training phase.

Furthermore, it is worth noticing that the segmentation information, which is produced as a side effect of spatial weighting, may be used for localization. It has been shown recently that combining the power of generative modeling with a discriminative classifier allows to obtain good results for object category localization [7]. We show promising results for the generation of segmentation masks for the Graz02 data set [18]. The results indicate that using the generated masks to guide the discriminative classifier can lead to the construction of a novel, effective localization method. For now, we show some preliminary results on the direct use of the masks for localization.

A somewhat similar solution for approximate image segmentation using local features and spatial information was proposed by Leibe et al. [13]. We have, however, experienced that a patch-based approach for segmentation does not work well in combination with a sparse image representation. The discriminative object parts that agree on the localization of the object can be covered with accurate segmentation patches. Nevertheless, for a full segmentation of the object, segmentation of non-discriminative object parts

Figure 1. PASCAL challenge image examples with ground-truth object annotation.

is also required. We find that a combination of full masks leads to better results compared to the use of local segmentation patches that we have implemented at an earlier stage of our research.

We avoided using the Hough transform [15], as it requires defining a parameter space and thus limits the shape hypotheses to forms like rectangles, ellipses, etc. Our method does not make assumptions about the object shape nor simplifies it to basic shapes. It uses the full shape information provided by the training objects and performs voting on the entire object boundary. For a given test image we compute a full segmentation mask estimate and obtain good results even for objects with irregular shapes.

In section 2 we describe the category classification framework of Zhang et al. [25] which is the bag-of-features approach our method builds on. We present our spatial weighting method in section 3. Experimental results are given in section 4. Proposed future work, including the discussion about using the method for object localization, is outlined in section 5.

2. Local features and kernels for object category recognition

In the following we describe the basic blocks of the framework by Zhang et al. [25] that we have extended. This method has shown excellent results in the PASCAL VOC Challenge [4] achieving the best classification accuracy for the more difficult test set. It first extracts an invariant image representation based on local image description and bag-of-features, and then uses non-linear Support Vector Machines

(SVMs) with extended Gaussian kernels for classification.

2.1. Detection of interest points

We use two complementary local region detectors to extract salient image structures: the Harris-Laplace detector [17] responding to corner-like regions and the Laplacian detector [14] extracting blob-like regions.

These two detectors are invariant to scale transformations, i.e., they output circular regions at a certain characteristic scale. To achieve rotation invariance, we may rotate the circular regions in the direction of the dominant gradient orientation [15, 17]. The affine adaptation procedure [8, 17] allows to obtain an affine-invariant version of the detectors. Affinely adapted detectors output elliptical regions which are then normalized into circles.

Note that it is unreasonable to use a more invariant description than required for a given data set [25]. For most natural object data sets the vertical direction is well defined, and the orientation of the features therefore contains valuable information. Thus, even though we can generalize our method to work with affinely adapted features, we will consider only the scale-invariant versions of the detectors in the experimental section. We will denote the scale-invariant Harris-Laplace detector as HS and the scale-invariant Laplacian detector as LS.

2.2. Local description

To compute appearance-based descriptors on the patches obtained by the detectors described in the previous subsection, we employ the SIFT [15] descriptor. We have also evaluated the SPIN [12] descriptor, but have not included

it into our final system, as it did not produce promising results.

The SIFT descriptor computes a gradient orientation histogram within the support region. For each of 8 orientation planes, the gradient image is sampled over a 4×4 grid of locations, thus resulting in a 128-dimensional feature vector for each region. A Gaussian window function is used to assign a weight to the magnitude of each sample point. This makes the descriptor less sensitive to small changes in the position of the support region and puts more emphasis on the gradients that are near the center of the region. To obtain robustness to illumination changes, the descriptors are made invariant to illumination transformations of the form $aI(x)+b$ by scaling the norm of each descriptor to unity [15].

Following the terminology of [12], we consider each detector/descriptor chain as a separate channel. We will denote the channels as HS-SIFT or LS-SIFT.

2.3. Bag-of-features representation

Given a set of local invariant descriptors, we want to represent their per-image distributions in training and test images. We therefore build a visual vocabulary by clustering the descriptors from the training set and then represent each image in the data set as a histogram of visual words drawn from the vocabulary [24]. Each histogram entry $h_{ij} \in H_i$ is the proportion of all descriptors in image i having label j to the total number of descriptors computed for the image.

Our evaluation has shown that vocabulary construction has little impact on the final classification results. We therefore randomly subsample the training set and cluster 50k features using K-means as clustering method to create a 1000-elements vocabulary.

2.4. Classification with non-linear SVMs

For classification, we use non-linear Support Vector Machines (SVMs) [22]. In a two-class setup that we use for binary detection, i.e., classifying images as containing or not containing a given object class, the decision function for a test sample x has the following form:

$$g(x) = \sum_i \alpha_i y_i K(x_i, x) - b \qquad (1)$$

where $K(x_i, x)$ is the value of a kernel function for the training sample x_i and the test sample x, $y_i \in \{+1, -1\}$ is the class label of x_i, α_i is a learned weight of the training sample x_i, and b is a learned threshold. The training samples with weight $\alpha_i > 0$ are usually called support vectors.

To obtain a detector response, we use the raw output of the SVM, given by eq. (1). By placing different thresholds on this output, we influence the decision and obtain Receiver Operating Characteristic (ROC) curves.

We use an extended Gaussian kernel [2, 11]:

$$K(H_i, H_j) = e^{-\frac{1}{A} D(H_i, H_j)} \qquad (2)$$

where $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$ are image histograms and $D(H_i, H_j)$ is the χ^2 distance defined as

$$D(H_i, H_j) = \frac{1}{2} \sum_{n=1}^{N} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}} \qquad (3)$$

where N is the size of the vocabulary ($N = 1000$ in our experiments). The resulting χ^2 kernel is a Mercer kernel [6]. The parameter A is the mean value of the distances between all training images [25].

We may combine different channels by summing their distances, so that $D = \sum_n D_n$ where D_n is the χ^2 distance for channel n. We will denote a combination of HS-SIFT and LS-SIFT channels as (HS+LS)-SIFT.

3. Spatial weighting

The idea of spatial weighting is to reduce the influence of background clutter by employing spatial relationships between the features. In the standard bag-of-features approach presented in section 2, each feature equally influences the bag-of-features representation. The goal of spatial weighting is to give lower weights to background features. This is achieved by having each feature boost other features that, from its spatial point of view, should belong to an object, e.g., a feature belonging to the wheel of a car should increase the weights of the features belonging to the other parts of the car.

See the fig. 2 for a visualization of the effect we want to achieve. Let's assume that the "drop" feature indicates the
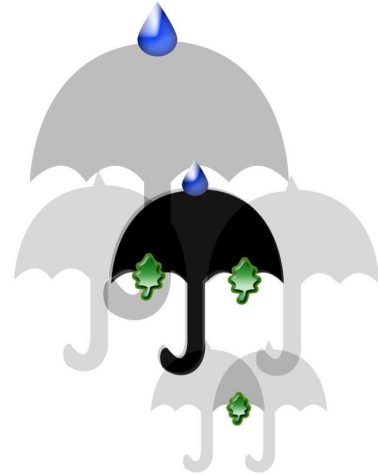

Figure 2. Visualization of spatial weighting. Three central features agree on the object localization, the two others are mistakes. Note the ambiguity introduced by the 'leaf' feature.

presence of the umbrella just below, while the "leaf" feature suggests that it is just above, but (we should consider possible ambiguity) it may be to the left or to the right, as leaves happen to stick to both sides of the umbrella. Given the set of features found on the test image, each feature may produce a hypothesis about the localization of the object. Note that true foreground features which agree on the position and shape of the object will quickly produce a strong response (the umbrella is obviously in the center of the visualization image) and those true foreground features will be rewarded later, as they are localized on the produced mask. The features that belong to the background clutter will not produce strong masks and will thus have low weights.

3.1. Potential of the approach

As we have described earlier, we have a strong motivation to employ spatial relationships between the features in the bag-of-keypoints representation. We can use the spatial information to estimate the position of the object in the image and discard the background. The usefulness of spatial weighting in the bag-of-features framework can be evaluated by using ground-truth segmentation. Fig. 4 shows that if we remove background clutter by using only foreground segments, we are able to significantly improve the classification results. The ROC curve achieved by testing on the object cropped out from an image is often well above the original ROC curve where testing is performed on the full image. Note, however, that training is always performed on full images, as it should not be performed using a training set that is easier than the expected test set, as was shown in [25].

3.2. Algorithm

In the following we explain how to produce a segmentation mask based on the training information. We will describe the use of the mask for background clutter reduction by boosting the foreground features and suppressing the background ones. For the application to localization, one should refer to subsection 5.2. Our spatial weighting procedure is described by the pseudo-code presented in listing 1.

During training, ground-truth segmentation information is used to learn (remember) the position of the object from a "point of view" relative to the training features. In fact, as we have ground-truth data for the training set, we first filter (line 1) the training data to include foreground features only. We perform this operation to avoid noise that would be introduced by matching test features with background features. The position of background features cannot be correlated with the position of the object, even if those features would give hints about the object category, e.g., a street sign could give us hints about cars object category, but it is impossible to draw any precise conclusions

```
01  TrainingSet.FilterFeatures();
02  for each TestImage in TestSet
03    Segmentation = 0;
04    for each TestFeature in TestImage
05      Hypothesis = 0;
06      for N closest TrainFeature in TrainingSet
07        M  = TrainFeature.GetImage().
              GetGroundTruthSegmentationMask();
08        T  = find_transformation(
              TrainFeature.GetPointOfView(),
              TestFeature.GetPointOfView());
09        M' = M.ApplyTransformation(T);
10        W  = gaussian(
              distance(TestFeature, TrainFeature),
              0, Sigma);
11        Hypothesis = Hypothesis + W * M';
12      Hypothesis.Normalize();
13      Segmentation = Segmentation + Hypothesis;
14    Histogram[TestImage] = 0;
15    for each TestFeature in TestImage
16      TestFeature.Weight(Segmentation);
17      Histogram[TestImage].Add(TestFeature);
```

Listing 1. Pseudo-code describing the spatial weighting procedure.

about the location of a car from the position of the sign. However, as segmentation information only roughly follows object edges and local descriptors need some support area, it is worth dilating (our choice) or blurring the segmentation image by some pixels (we chose 32) before filtering out the background features.

Having prepared the training set, we can generate hypotheses about possible object locations and shapes for each feature of a test image (line 4). For each test feature we look for the features from the training data that are closest (we use Euclidean distance here) in the 128-dimensional feature-space (for our SIFT implementation). We choose $N = 100$ most similar training features (line 6). For each interest point that is found, we are given not only its position, but with scale-invariant detectors we also know its scale [14]. By finding dominant gradient orientation [15] we may determine the orientation of the point and with affine-adaptation technique [8, 17] it is possible to find all affine deformation parameters. We call this information a "point of view" of a given feature. The point of view is necessary to normalize the retrieved mask shapes (line 7) to compensate for viewpoint changes (lines 8-9). For example, in the case of scale invariant features, a feature detected at scale 6 may correspond to a feature detected at scale 3 in the training image. Then we need to shift the mask of the training image to the relative position of the point in the test image and rescale the mask by a factor of 2. In the same manner we use rotation compensation for rotation invariant features and affine transformation for affine invariant features. We sum the transformed masks and create a hypothesis cast by the test feature (line 11). Masks in the

		Winner [4]	Reimpl. of Zhang et al. [25]			Spatial weighting			Gain
			HS--SIFT	LS--SIFT	(HS+LS)--SIFT	HS--SIFT	LS--SIFT	(HS+LS)--SIFT	
test set 1	bikes	**93.0**	85.1	90.4	92.1	86.8	91.2	92.1	
	cars	**96.1**	93.5	93.8	94.5	93.5	94.9	**96.0**	+1.5
	motorbikes	**97.7**	94.0	95.8	96.3	92.6	95.4	96.3	
	people	91.7	89.3	88.1	91.7	89.3	89.3	**92.9**	+1.2
test set 2	bikes	72.8	72.6	73.4	74.8	75.3	75.9	**76.8**	+2.0
	cars	72.0	72.5	73.9	75.8	73.7	73.9	**76.8**	+1.0
	motorbikes	**79.8**	72.9	77.1	78.8	74.3	78.2	79.3	+0.5
	people	71.9	75.1	74.5	76.9	76.3	74.9	**77.9**	+1.0

Table 1. Equal Error Rates (EER) of ROC curves for the classification task of the PASCAL challenge. Best result achieved during the challenge ('winner'), performance of our reimplementation of Zhang's method and improvement introduced with our spatial weighting are presented.

sum are weighted with a Gaussian function of the distance between the training and test features (line 10). We have found $Sigma = 0.15$ to be a reasonable value for σ.

The normalized hypotheses (line 12) of all the features in the test image are added to create the segmentation mask (line 13), see fig. 3 for examples of the resulting masks.

Figure 3. Test images of Graz02 data set (on the left), generated masks (in the middle) and multiplication of the two (on the right).

One may also consider the final mask to be a score map describing the likelihood that a given image pixel belongs to an object. The score value is then computed based on the number of features agreeing on a hypothesis that a given pixel is a foreground pixel, i.e., belongs to an object. The scores are then used to boost the importance of the features lying on the object and suppress the background features. We have redesigned the histogram building algorithm to weight features according to the mask value corresponding to their positions (line 16). Thus the features considered to be foreground features will most strongly influence the rep-

resentation and the background features will have a minor impact on the histogram.

4. Experimental results

To measure the performance of our spatial weighting technique, we have reimplemented the method of Zhang et al. [25] and evaluated it on the PASCAL VOC Challenge [4] data set with and without spatial weighting. We have evaluated the binary classification performance using Receiver Operating Characteristic (ROC). We have performed quantitative evaluation of the ROC curves by computing the Equal Error Rates (EER) following the procedure defined for the challenge [4].

Table 1 summarizes the results. For each of the eight test sets we present the best reported result of the challenge, the performance of the evaluated method without spatial weighting (for each channel separately and for the combination) and the performance after introducing our technique (again for each channel separately and for the combination). We also show the gain achieved by our method.

One may have the impression that the achieved improvement is not so high compared to the overall result. However, our 1.5% improvement should not be compared to the original 94.5% ERR, but rather to the remaining 5.5%. It should be taken into account that we are improving the method, which gives the best known results for 5 out of 8 PASCAL test sets and for all of the most challenging ones. Our method achieves the similar performance on the easier datasets and outperforms the best known methods on 3 out of 4 more difficult test sets.

It is worth noting that the performance of the reimplemented method itself is slightly better than the one reported during the challenge [4]. This is due to the improvement in the A parameter selection [25] for the χ^2 kernel (see eq. (2) in subsection 2.4).
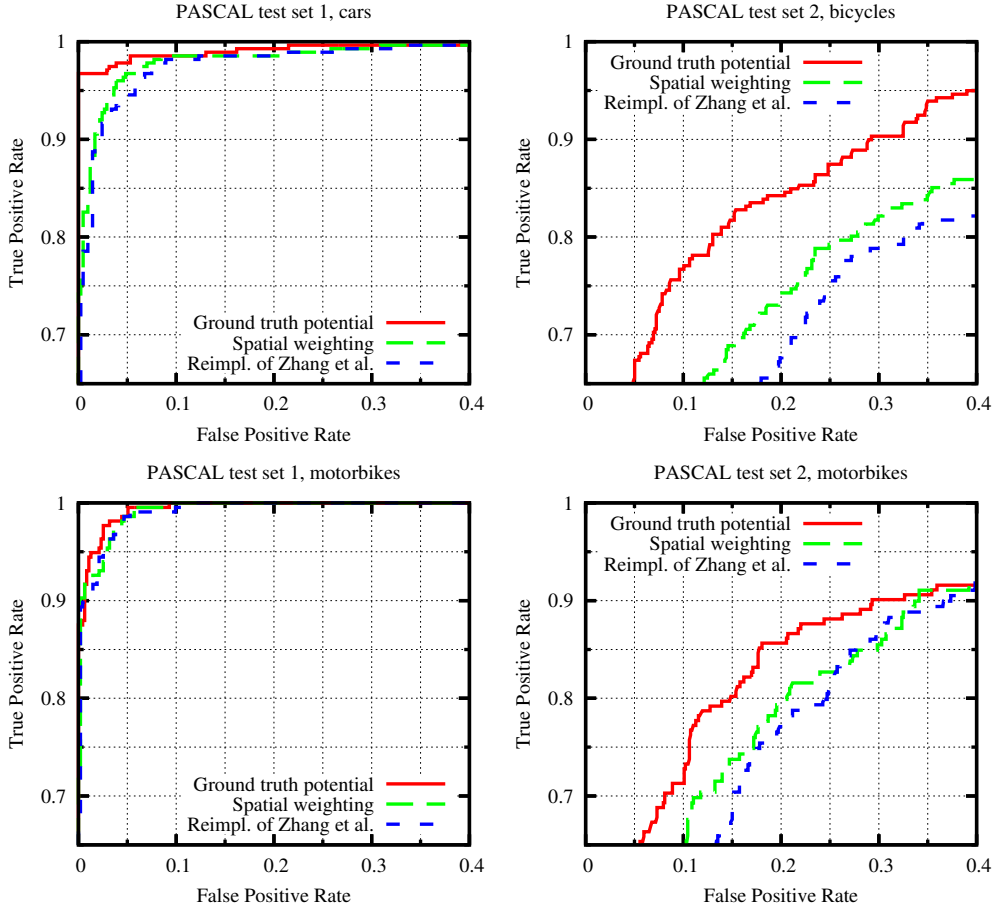
Figure 4. Selected Receiver Operating Characteristic (ROC) curves for classification on PASCAL test sets. Results for a reimplementation of Zhang et al. [25] framework, our spatial weighting method and ground-truth based segmentation are presented. Please note that only the most interesting parts of the curves are shown to improve readability.

Fig. 4 presents some of the computed ROC curves. Plots are shown for the evaluated method without spatial weighting, for the same method with spatial weighting and for a ground-truth estimated potential that motivated us to develop the technique. The curves were selected to show cases with and without improvement. Results for other test sets are supporting our conclusions, the selection was done purely due to space limitations.

One can notice that the method does not give improvement for all test sets. It is worth comparing the achieved gain with the potential estimated using the ground-truth segmentation. A conclusion may be drawn that in cases where the ground-truth ROC curve falls close to the ROC curve of the original approach, minor improvement or no improvement at all is observed. This can be easily understood. If there is little background clutter, removal of background does not help. For example, the background of motorbikes in test set 1 is mostly uniform and the EER cannot be pushed above 96.3%. On the other hand, the background of bicy-

cles in test set 2 is heavily cluttered and spatial weighting gives 2.0% improvement in this case.

5. Extensions and future work

Spatial weighting has several potential extensions. Two of them are discussed in the following. We also show some promising initial results for localization.

5.1. Feature selection and iterative version

In the same way as we filter the training feature set to reduce the noise caused by false matches with background features, we could try to filter the test image features to reduce the number of considered background features. Filtering with ground-truth segmentation information reveals a potential of improving the results presented in section 4 by further 1% on average and even by up to 2.5% in the case of Pacal test set 1 containing people. Naturally, we may not use ground-truth for the purpose described above,
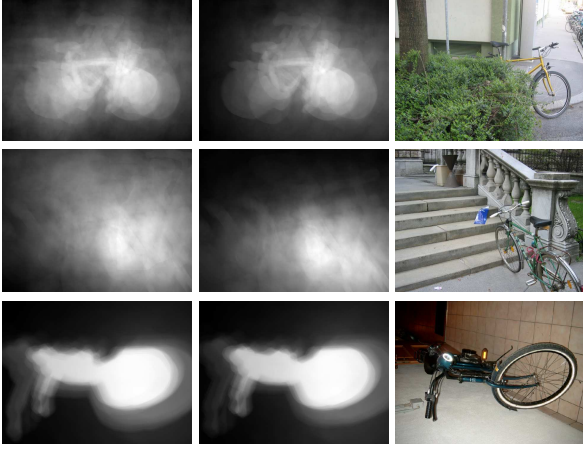
Figure 5. Iterating the spatial weighting procedure: first iteration (on the left), last iteration (in the middle) and original image (on the right).

	Opelt [20]	Spatial weighting	
		HS-SIFT	LS-SIFT
bikes	76.7	78.7	**82.7**
cars	55.3	62.7	**68.0**
people	48.0	**83.3**	71.3

Table 2. Percentage of the images that satisfy the localization criterion [20].

so one could choose to implement feature selection techniques, e.g., the likelihood ratio of the classification [3].

We have decided to evaluate an iterative approach to spatial weighting instead. It is possible to build segmentation masks iteratively, by weighting the test features using the segmentation mask from the previous step. Effectively, we should achieve a result similar to filtering with ground-truth information. The masks usually converge to a stable form in about 10 iterations. See fig. 5 for overview of the effect. Note the difference between more (top two rows) and less (bottom row) cluttered images. From top to bottom it took 9, 8 and 3 iterations to converge.

5.2. Localization

Segmentation masks produced by the spatial weighting method seem to be promising for localization. Note that they are not sufficient for the general localization task, where we have to distinguish between separate object instances on one image. However, the masks are directly suitable for a task where we can assume that one object per image is present. Here we search for the highest value in the mask to select the point with the highest probability of being localized on the object. Object boundary approximation can be determined by simple thresholding. We plan to further develop the method to support multiple object instances on one image.

As the PASCAL data set contains multiple object instances per image, it is not suitable for evaluation of the produced segmentation masks. We have therefore followed the experimental setting of Opelt and Pinz [20] defined for the Graz02 [18] dataset. We have evaluated the localization performance of our method following the criterion chosen by the authors. It is based on the criterion established

by Agarwal et al. [1], which requires the position given by the system to fall within an ellipse drawn at the center of the localized object. However, due to different parameter settings, Opelt's ellipse is larger and thus the criterion is weaker then Agarwal's. Table 2 presents the comparison with Opelt's approach.

We have also evaluated the localization accuracy by generalizing the bounding box evaluation criterion defined for the PASCAL challenge [4] to any bounding shapes. We varied the overlap requirement, i.e., the localization was considered to be correct when

$$\frac{|H \cap G|}{|H \cup G|} \geq t \qquad (4)$$

where H is a computed localization mask, G is ground-truth localization mask and t is an overlap threshold. Fig. 6 shows the localization accuracy—i.e, the number of correct localizations with respect to the total number of test images—as a function of the threshold t. As one could expect, our method gives more precise localization for rigid objects like cars and fails where very precise segmentation is required for objects with highly variable silhouettes like people. The localization accuracy grows rapidly while relaxing the threshold. Please note that for arbitrary silhouettes the masks fulfilling the 33% overlap requirement are usually visually satisfying. We only give results for the de-
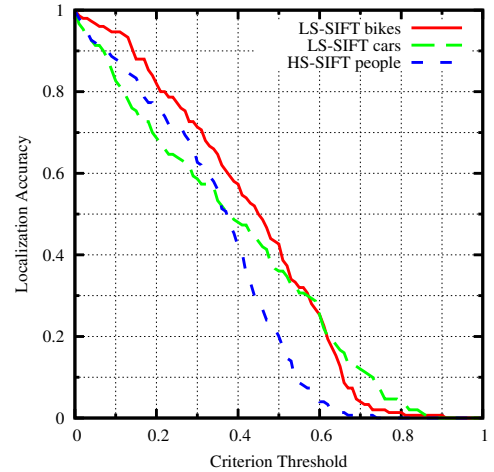

Figure 6. Localization accuracy evaluated using the overlap criterion.

tector that obtained best results in table 2. The other detector achieved slightly lower results as expected.

6. Summary

In this paper we have proposed an extension to category classification with bag-of-features that incorporates spatial relations between features. We have introduced the "spatial weighting" technique, which uses spatial relations to boost the weights of foreground features and to decrease the influence of background features on the representation, thus making it more robust to background clutter.

The experimental evaluation has shown that applying the proposed extension to one of the state-of-the-art methods further improves the classification results. The classification rate achieved by our method on the PASCAL VOC Challenge data set outperforms the state-of-the-art [4].

We have also demonstrated the possibility of applying our method to object localization. Preliminary results show promise. Future research could focus on guiding a discriminative classifier using the maps produced by the spatial weighting technique to create an efficient localization method.

Acknowledgments

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11), 2004.

[2] O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *NN*, 10(5):1055–1064, 1999.

[3] G. Dorkó and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, volume 1, pages 634–640, 2003.

[4] M. Everingham, A. Zisserman, C. Williams, L. V. Gool, et al. The 2005 PASCAL visual object classes challenge. In *First PASCAL Challenge Workshop*.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, 2003.

[6] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *PAMI*, 26(2):1–12, 2004.

[7] M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminant models for object category detection. In *ICCV*, volume 2, pages 1363–1370, 2005.

[8] J. Gårding and T. Lindeberg. Direct computation of shape cues using scale-adapted spatial derivative operators. *IJCV*, 17(2):163–191, 1996.

[9] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, volume 2, pages 1458–1465, 2005.

[10] E. Hayman, B. Caputo, M. Fritz, and J.-O. Eklundh. On the significance of real-world conditions for material classification. In *ECCV*, pages 253–266, 2004.

[11] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. Support vector machines for region-based image retrieval. In *ICME*, 2003.

[12] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *ICCV*, volume 1, pages 832–838, 2005.

[13] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV'04 Workshop on Statistical Learning in Computer Vision*, pages 17–32, 2004.

[14] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.

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

[16] S. Lyu. Mercer kernels for object recognition with local features. In *CVPR*, volume 2, pages 223–229, 2005.

[17] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[18] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Generic object recognition with boosting. Technical Report TR-EMT-2004-01, TU Graz, 2004.

[19] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV*, volume 2, pages 71–84, 2004.

[20] A. Opelt and A. Pinz. Object localization with boosting and weak supervision for generic object recognition. In *SCIA*, 2005.

[21] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000.

[22] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA, 2002.

[23] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *ICCV*, volume 1, pages 370–377, 2005.

[24] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *IWLAVS*, 2004.

[25] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classifcation of texture and object categories: An in-depth study. Technical Report RR-5737, INRIA Rhône-Alpes, Nov 2005.