# Object Recognition and Localization via Spatial Instance Embedding

Nazli Ikizler-Cinbis and Stan Sclaroff
*Boston University, Department of Computer Science*
*Boston, MA, USA*
{*ncinbis,sclaroff*}*@cs.bu.edu*

***Abstract*—We propose an approach for improving object recognition and localization using spatial kernels together with instance embedding. Our approach treats each image as a bag of instances (image features) within a multiple instance learning framework, where the relative locations of the instances are considered as well as the appearance similarity of the localized image features. The introduced spatial kernel augments the recognition power of the instance embedding in an intuitive and effective way, providing increased localization performance. We test our approach over two object datasets and present promising results.**

***Keywords*-object recognition; object localization; multiple instance learning;**

## I. Introduction

Object recognition and localization are two major problems in computer vision. For the object recognition problem, bag-of-words approaches [1], [2] have recently gained a lot of interest in the community, due to their simplicity and effectiveness. In such approaches, extracting local features and representing the image with the histogram of these local features is a common practice. However, bag-of-words approaches have certain shortcomings. First, using pure histogramming over the image ignores the important spatial information present in the 2D image domain. Second, hard assignment of interest points to codewords is prone to noise caused by background features.

Using localized features, the problem of object recognition and localization can be formulated as a *multiple instance learning (MIL)* problem, where the image features/regions represent the instances and the whole image or a subwindow can be considered as a bag. Then, the problem reduces to finding the correct set of instances, i.e. features, that represent a particular class. Following the instance embedding approach of Chen, et al. [3], we can define a mapping so that each image is represented by the overall distances of its regions to a global dictionary of localized features. This approach overcomes the shortcomings of the bag-of-words approach such that: 1) using interest points as is, the overhead introduced by the codebook generation step is eliminated, and 2) each image is represented in terms of a dictionary, which provides a higher level of tolerance for noisy features. This approach is powerful in finding the relevant patches in images. However, in the 2D image domain, the spatial layout of the image patches is also important. Therefore, we propose to add spatial reasoning to the formulation of instance embedding by means of a spatial kernel. In this way, we aim to achieve better localization and recognition. Moreover, this spatial information is likely to improve the instance selection process of the MIL problem.

Some approaches have looked at exploiting spatial information by means of spatial binning [4], spatial pyramid histograms [5], generalized Hough transform [6], [7]. None of these approaches has formulated the problem in a multiple instance learning (MIL) framework. In this paper we look at how we may improve over the current solutions by incorporating this spatial information. We achieve this by formulating a spatial kernel, which is easily compatible with the multiple instance embedding approach of [3].

We evaluate both the object recognition and localization performance of our proposed algorithm, using the Caltech-4, the UIUC multi-scale cars and Graz-02 datasets. The results show that our approach is successful in both recognition and localization of the objects. In these experiments, we show that spatial reasoning provides more successful localization for the instance embedding approach, and the results compare favorably to various methods presented in the literature [8], [9], [10], [11].

## II. Our Approach

Our approach is built upon the localized features within the image and is an alternative to the bag-of-words representation. The regular bag-of-words approach first generates a codebook by clustering the image patches. Then, each image is represented with a subset of this codebook, such that each image patch is represented with the closest codeword. Then, the overall image is represented using a histogram of these codewords and all the image contents are accumulated into bins.

There are several shortcomings of this approach. First, the codebook generation can be imperfect. Once the codebook is formed, hard assignment of the interest points to the closest cluster centers, i.e. assigning each patch to the closest codebook entry, may cause information loss. That is why some approaches use soft-assignment [4], rather than using the closest cluster center.

Following [3], an image can be represented by not only the closest codewords, but in terms of all the dictionary. A discriminative classifier can then be used to select the

important features and dictionary points. In this setting,we define an instance embedding space such that, given the entire instance space or codebook $C = \{c^1, c^2, \ldots, c^N\}$, we represent each image $i$ with embedded feature vector $\mathbf{m}(B_i) = \{s(B_i, c^1), s(B_i, c^2), \ldots, s(B_i, c^N)\}$, where $s(B_i, c^k)$ represents the similarity between the image, and the codeword $c^k$ in the dictionary. In this way, we convert the input data vector to its alternative representation in the space of the codebook dictionary. By using the exact distances to all codebook entries, the pitfalls of the hard assignment are avoided. In [3], $s$ is formulated as follows:

$$s(\mathbf{B}_i, c^k) = \max_j \exp\left(-\frac{\left\|x_{ij} - c^k\right\|^2}{\sigma^2}\right) \quad (1)$$

where $x_{ij}$ represents the $j$th feature vector for image $i$.

This is a multiple instance learning (MIL) formulation where the task is to select "correct" instances towards learning a good model. This embedded space converts the instance selection problem into a feature selection problem. The noisy instances are expected to be inconsistent in the dataset, and in this embedding, they will appear as less informative feature dimensions.

In this image-based MIL setting, spatial locations of instances (interest points) can provide strong additional information that can be considered as a prior to select the positive instances in each image. Fortunately, it is straightforward to add such spatial reasoning to this MIL framework. We introduce a multiplicative spatial kernel to the feature-based similarity measure, and represent $s$ as follows:

$$s(\mathbf{B}_i, c^k) = \max_j \left(\phi_{feat}(x_{ij}, c^k)\phi_{spatial}(x_{ij}, c^k)\right) \quad (2)$$

where $\phi_{feat}$ is the similarity between feature vectors of instances and $\phi_{spatial}$ is the spatial closeness. $\phi_{feat}$ and $\phi_{spatial}$ are defined as

$$\phi_{feat}(x_{ij}, c^k) = \exp\left(-\frac{D(x_{ij}, c^k)}{\sigma^2_{feat}}\right), \quad (3)$$

$$\phi_{spatial}(x_{ij}, c^k) = \exp\left(-\frac{\left\|P(x_{ij}) - P(c^k)\right\|^2}{\sigma^2_{spatial}}\right). \quad (4)$$

In Eq. 3, $D$ corresponds to the distance measure used to compute the similarity of two feature vectors, and the choice of $D$ depends on the application. In our case, we use $D(x_{ij}, c^k) = \chi^2(x_{ij}, c^k) = \frac{1}{2}\sum_n \frac{(x_{ij}(n) - c^k(n))^2}{x_{ij}(n) + c^k(n)}$ to compute the similarity of two SIFT vectors.

In Eq. 4, the spatial positions $P(x)$ of the image patches are compared and their distances are encoded within the final similarity measure. This extended formulation allows us to consider the relative spatial locations of the feature vectors as well as their content similarity. This is achieved by using the direct Euclidean distance between feature locations, without any spatial binning. Each feature location is normalized with respect to the search window size, in order to achieve invariance to differences in scale. The scalars $\sigma_{feat}$ and $\sigma_{spatial}$ are predefined bandwidth parameters that are used to scale each of the distance kernels. The bandwidth parameter helps to adjust the sensitivity of the measure to the spatial differences. These parameters can be selected by using cross-validation over the training set.

We then apply L1-regularized linear SVM over this instance embedding representation as in [3]. L1 regularization SVM provides us implicit feature selection, so that in the test phase, we only use those instances that have non-zero weights. In this case, L1 SVM associates a weight $w_j$ with each instance, by minimizing

$$||\mathbf{w}||_1 + C\sum_i L(f(\mathbf{m}(B_i)), y_i) \quad (5)$$

where L is the hinge loss. $f(B_i)$ is the linear classification function defined as

$$f(\mathbf{m}(B_i)) = \sum_k w_k s(B_i, c^k) + b. \quad (6)$$

Table I
TRUE POSITIVE RATES AT THE EQUAL ERROR RATE POINT(EER) ON CALTECH-4 DATASET.

| Approach | Airplanes | Cars | Faces | Mbikes |
|---|---|---|---|---|
| spMILES | **98.25** | **93.25** | 99.08 | **98.75** |
| MILES [3] | 96.0 | 89.75 | **99.54** | 97.75 |

## III. EXPERIMENTS

We have conducted two sets of experiments: the first one is to measure the object recognition performance, and the second one is to measure the object localization performance.

### A. Object Recognition

For the object class recognition problem, we use the Caltech dataset from [8]. This dataset contains four object classes, namely airplanes, cars, faces and motorbikes, together with the background images. We extracted 128-dimensional SIFT features [12] from the interest point regions detected by the Harris-Hessian-Laplace [13] interest point detector. We used the same set of features both for MILES [3] and our spatial MILES (spMILES) approach.

The whole instance set $x^k$ in the training phase yields $\approx$ 30000 instances. We can either use the whole set or a random subset of instances, or apply an initial clustering to reduce the instance space. In our preliminary experiments, we have not seen any significant performance difference between using the whole set versus using the cluster centers; thus, we chose to cluster the instances first (using k-means with $k = 3000$) and use the instances that are closest to the cluster centers as our reduced instance space. Note that this clustering step is not as critical as in quantization-based approaches. Here, clustering is used only as a well-defined

Table II
TRUE POSITIVE RATES AT THE EQUAL ERROR RATE POINT(EER) ON CALTECH-4 DATASET.

| Approach | Airplanes | Cars | Faces | Mbikes |
|---|---|---|---|---|
| spMILES | **98.25** | 93.25 | 99.08 | **98.75** |
| Fergus [8] | 90.2 | 90.3 | 96.4 | 92.5 |
| Opelt [11] | 88.9 | 90.1 | 93.5 | 92.2 |
| Loeff [9] | 97.0 | 98.0 | 98.7 | 97.0 |
| Bar-Hillel [10] | 93.3 | **99.4** | 93.7 | 95.1 |

procedure to reduce the number of features to a computationally feasible subset, rather than building a quantization codebook. We observe that L1-regularized SVM provides good generalization with selecting as few as $\sim 200$ instances in the final model.

Table I shows the object recognition performance of the proposed approach (spMILES). To make direct comparison possible, we evaluate MILES and spMILES over the same set of SIFT features, which are extracted as described above. As can be seen, the spatial kernel helps in improving the recognition rates in most of the classes.

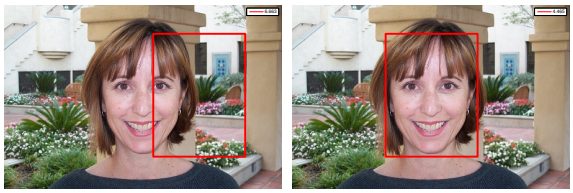
Table II shows the recognition performance relative to the other methods proposed. The recognition rate of spMILES is quite competent. We should note that the methods presented in this table are not directly comparable, because they operate over different feature sets. Our method possibly uses sub-optimal features. Nevertheless, our main point is to demonstrate the improvement that is possible when instance embedding is done with spatial reasoning. Various studies have shown that optimized or multiple sets of features may yield further performance improvement and this remains as a topic for future work.

### B. Object Localization

A main strength of the proposed approach is in its power to localize object instances. Figure 1 shows examples of localization. While MILES is quite powerful in the binary decision about the presence of the object of interest, it is not quite good at localization. By adding spatial reasoning in the form of the spatial kernel, spMILES correctly localizes the object. We perform localization by using a sliding window approach over multiple scales. Candidate subwindows are evaluated with respect to their SVM output over the spatial embedding domain.

For evaluating object localization, we use the UIUC multi-scale cars dataset [14] which consists of images of cars at multiple scales and in multiple locations. There can be more than one car in an image and there can be some occlusions. Figure 2 shows example car detections in this dataset and Table III shows the comparison of the localization performance. The average precision rates are calculated by ranking the positive detections by their output score and the detections that have more than $50\%$ overlap with the ground truth locations are considered to be true positives.

(a) airplanes



(b) faces

(c) motorbikes

Figure 1. Spatial reasoning helps multiple instance learning and improves localization of the objects.

Table III
AVERAGE PRECISION (AP) RATES FOR OBJECT LOCALIZATION EXPERIMENT ON UIUC CARS DATASET.

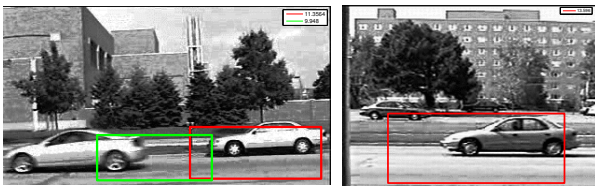
| | MILES | spMILES | [2] |
|---|---|---|---|
| localization AP | 19.17 | 90.3 | 90.6 |

As can be seen, although MILES has high recognition rates, without the spatial reasoning, it tends to yield incorrect localization. This situation can also be observed from the example images given in Fig. 2.

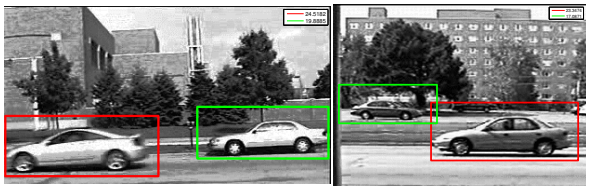
Figure 3 shows some localization results from the more challenging Graz-02 dataset [11]. In this dataset there are severe occlusions, as well as viewpoint and scale changes. In order to compensate for viewpoint changes, we apply sliding window technique over multiple aspect ratios (i.e. 0.5,1,1.5). As seen from the examples in Fig 3, the spMILES approach is able to detect the object of interest successfully in many difficult cases.

## IV. CONCLUSION

In this paper, we present a multiple instance learning(MIL)-based approach for object recognition and localization. Our approach extends the discriminative MIL framework to image domain by using spatial information by means of a spatial kernel. Our formulation is directly compatible with the instance embedding framework introduced in [3]. The results demonstrate that the proposed



(a) MILES output



(b) spMILES output

Figure 2. Localization examples from UIUC multi-scale cars dataset.

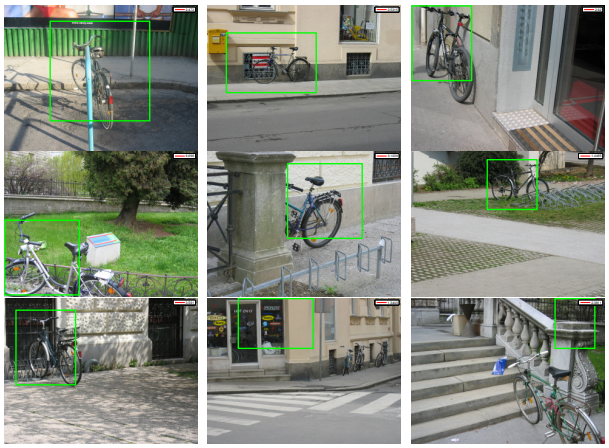


Figure 3. Localization examples for bicycle class from Graz-02 dataset. We perform a sliding-window search over multiple scales and aspect ratios to accommodate for differences in orientations of objects.

approach offers considerable improvement over the object recognition and localization performance as compared to using multiple instance embedding [3] alone.

## ACKNOWLEDGMENT

This material is based upon work supported in part by the U.S. National Science Foundation under Grant No. 0713168.

## REFERENCES

[1] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *Proceedings of the International Conference on Computer Vision*, 2005.

[2] J. Mutch and D. Lowe, "Multiclass object recognition with sparse, localized features," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[3] Y. Chen, J. Bi, and J. Z. Wang, "Miles: Multiple-instance learning via embedded instance selection," *PAMI*, vol. 28, pp. 1931–1947, 2006.

[4] T. Quack, V. Ferrari, B. Leibe, and L. V. Gool, "Efficient mining of frequent and distinctive feature configurations," in *Int. Conf. on Computer Vision*, 2007.

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

[6] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *IJCV*, vol. 77, no. 1, pp. 259–289, 2008.

[7] S. Maji and J. Malik, "Object detection using a max-margin hough transform," in *CVPR*, 2009.

[8] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[9] N. Loeff, A. Sorokin, and D. A. Forsyth, "Efficient unsupervised learning for localization and detection in object categories," in *NIPS*, 2005.

[10] A. Bar-Hillel and D. Weinshall, "Efficient learning of relational object class models," *Int. J. Comput. Vision*, vol. 77, pp. 175–198, May 2008.

[11] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *PAMI*, vol. 28, no. 3, 2006.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1, pp. 43–72, 2005.

[14] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *PAMI*, pp. 1475–1490, 2004.