in *DAGM'04 Pattern Recognition Symposium*, Tübingen, Germany, Aug 2004.

Scale-Invariant Object Categorization using a Scale-Adaptive Mean-Shift Search

Bastian Leibe[1] and Bernt Schiele[1,2]

[1] Perceptual Computing and Computer Vision Group, ETH Zurich, Switzerland
leibe@inf.ethz.ch, http://www.vision.ethz.ch/leibe

[2] Multimodal Interactive Systems, TU Darmstadt, Germany
schiele@informatik.tu-darmstadt.de

Abstract. The goal of our work is object categorization in real-world scenes. That is, given a novel image we want to recognize and localize unseen-before objects based on their similarity to a learned object category. For use in a real-world system, it is important that this includes the ability to recognize objects at multiple scales.
In this paper, we present an approach to multi-scale object categorization using scale-invariant interest points and a scale-adaptive Mean-Shift search. The approach builds on the method from [12], which has been demonstrated to achieve excellent results for the single-scale case, and extends it to multiple scales. We present an experimental comparison of the influence of different interest point operators and quantitatively show the method's robustness to large scale changes.

1 Introduction

Many current object detection methods deal with the scale problem by performing an exhaustive search over all possible object positions and scales [17–19]. This exhaustive search imposes severe constraints, both on the detector's computational complexity and on its discriminance, since a large number of potential false positives need to be excluded. An opposite approach is to let the search be guided by image structures that give cues about the object scale. In such a system, an initial interest point detector tries to find structures whose extend can be reliably estimated under scale changes. These structures are then combined to derive a comparatively small number of hypotheses for object locations and scales. Only those hypotheses that pass an initial plausibility test need to be examined in detail. In recent years, a range of scale-invariant interest point detectors have become available which can be used for this purpose [13–15, 10].

In this paper, we apply this idea to extend the method from [12, 11]. This method has recently been demonstrated to yield excellent object detection results and high robustness to occlusions [11]. However, it has so far only been defined for categorizing objects at a known scale. In practical applications, this is almost never the case. Even in scenarios where the camera location is relatively fixed, objects of interest may still exhibit scale changes of at least a factor of two simply because they occur at different distances to the camera. Scale invariance is thus one of the most important properties for any system that shall be applied to real-world scenarios without human intervention.

This paper contains four main contributions: (1) We extend our approach from [12, 11] to multi-scale object categorization, making it thus usable in practice. Our extension is based on the use of scale-invariant interest point detectors, as motivated above. (2) We formulate the multi-scale object detection problem in a Mean-Shift framework, which allows to draw parallels to Parzen window probability density estimation. We

Method	Agarwal et al [1]	Garg et al [9]	Leibe et al [11]	Fergus et al [8]	Our approach
Equal Error Rate	∼79%	∼88%	97.5%	88.5%	91.0%
Scale Inv.	no	no	no	yes	yes

Table 1. *Comparison of results on the UIUC car database reported in the literature.*

show that the introduction of a scale dimension in this scheme requires the Mean-Shift approach to be extended by a scale adaption mechanism that is different from the variable-bandwidth methods proposed so far [6, 4]. (3) We experimentally evaluate the suitability of different scale-invariant interest point detectors and analyze their influence on the recognition results. Interest point detectors have so far mainly been evaluated in terms of repeatability and the ability to find exact correspondences [15, 16]. As our task requires the generalization to unseen objects, we are more interested in finding similar and typical structures, which imposes different constraints on the detectors. (4) Last but not least, we experimentally evaluate the robustness of the proposed approach to large scale changes. While other approaches have used multi-scale interest points also for object class recognition [7, 8], no quantitative analysis of their robustness to scale changes has been reported. Our results show that the proposed approach outperforms state-of-the-art methods while being robust to scale changes of more than a factor of two. In addition, our quantitative results allow to draw some interesting conclusions for the design of suitable interest point detectors.

The paper is structured as follows. The next section discusses related work. After that, we briefly review the original single-scale approach. Section 3 then describes our extension to multiple scales. In Section 4, we examine the influence of different interest point detectors on the recognition result. Finally, Section 5 evaluates the robustness to scale changes.

2 Related Work

Many current methods for detection and recognition of object classes learn global or local features in fixed configurations or using configuration classifiers [17–19]. They recognize objects of different sizes by performing an exhaustive search over scales. Other approaches represent objects by more flexible models involving hand-defined or learned object parts. [20] models the joint spatial probability distribution of such parts, but does not explicitly deal with scale changes. [8] extends this approach to learn scale-invariant object parts and estimates their joint spatial and appearance distribution. However, the complexity of this combined estimation step restricts the method to a small number of parts. [7] also describes a method for selecting scale-invariant object parts, but this method is currently defined only for part detection, not yet on an object level. Most directly related to our approach, [1] learns a vocabulary of object parts for recognition and applies a SNoW classifier on top of them (which is later combined with the output of a more global classifier in [9]). [3] learns a similar vocabulary for generating class-specific segmentations. Both approaches only consider objects at a single scale. Our approach combines both ideas and integrates the two processes of recognition and figure-ground segmentation into a common probabilistic framework [12, 11], which will also be the basis for our scale-invariant system. The following section briefly reviews this approach. As space does not permit to give a complete description, we only highlight the most important points and refer to [12, 11] for details.

2.1 Basic Approach

The variability of a given object category is represented by learning, in a first step, a class-specific codebook of local appearances. For this, fixed-size image patches are extracted around Harris interest points from a set of training images and are clustered with an agglomerative clustering scheme. We then learn the spatial distribution of codebook entries for the given category by storing all locations the codebook entries were matched to on the training objects. During recognition, this information is used in a probabilistic extension of the Generalized Hough Transform [2, 14]. Each patch **e** observed at location ℓ casts probabilistic votes for different object identities o_n and positions x according to the following equation:

$$p(o_n, x|\mathbf{e}, \ell) = \sum_i p(o_n, x|I_i, \ell)p(I_i|\mathbf{e}). \quad (1)$$

where $p(I_i|\mathbf{e})$ denotes the probability that patch **e** matches to codebook entry I_i, and $p(o_n, x|I_i, \ell)$ describes the stored spatial probability distribution for the object center relative to an occurrence of that codebook entry. In [12], object hypotheses are found as maxima in the voting space using a fixed-size search window W:

$$score(o_n, x) = \sum_k \sum_{x_j \in W(x)} p(o_n, x_j|\mathbf{e}_k, \ell_k). \quad (2)$$

For each such hypothesis, we then obtain the per-pixel probabilities of each pixel being *figure* or *ground* by the following double marginalization, thus effectively segmenting the object from the background (again see [12, 11] for details):

$$p(\mathbf{p} = \textit{figure}|o_n, x) = \sum_{\mathbf{p}\in(\mathbf{e},\ell)} \sum_I p(\mathbf{p} = \textit{fig.}|o_n, x, I, \ell) \frac{p(o_n, x|I, \ell)p(I|\mathbf{e})p(\mathbf{e}, \ell)}{p(o_n, x)} \quad (3)$$

The per-pixel probabilities are then used in an MDL-based hypothesis verification stage in order to integrate only information about the object and discard misleading influences from the background [11]. The resulting approach achieves impressive results (as a comparison with other methods in Tab. 1 shows), but it has the inherent limitation that it can only recognize objects at a known scale. In practical applications, however, the exact scale of objects is typically not known beforehand, and there may even be several objects with different scales in the same scene. In order to make the approach applicable in practice, it is thus necessary to achieve scale-invariant recognition.

3 Extension to Multiple Scales

A major point of this paper is to extend recognition to multiple scales using scale-invariant interest points. The basic idea behind this is to replace the single-scale Harris codebook used up to now by a codebook derived from a scale-invariant detector. Given an input image, the system applies the detector and obtains a vector of point locations, together with their associated scales. Patches are extracted around the detected locations with a radius relative to the scale σ of the interest point (here: $r = 3\sigma$). In order to match image structures at different scales, the patches are then rescaled to the codebook size (in our case 25×25 pixels).

The probabilistic framework can be readily extended to multiple scales by treating scale as a third dimension in the voting space. If an image patch found at location $(x_{img}, y_{img}, s_{img})$ matches to a codebook entry that has been observed at position $(x_{occ}, y_{occ}, s_{occ})$ on a training image, it votes for the following coordinates:

$$x_{vote} = x_{img} - x_{occ}(s_{img}/s_{occ}) \quad (4)$$

$$y_{vote} = y_{img} - y_{occ}(s_{img}/s_{occ}) \quad (5)$$

$$s_{vote} = (s_{img}/s_{occ}) \quad (6)$$

However, the increased dimension of the voting space makes the maxima search computationally more expensive. For this reason, we employ a two-stage search strategy. In a first stage, votes are collected in a binned 3D Hough accumulator array in order to quickly find local maxima. Candidate maxima from this first stage are then refined in the second stage using the original (continuous) 3D votes. Instead of a simple but expensive sliding-window technique, we formulate the search in a Mean-Shift framework. For this, we replace the simple search window W from equation (2) by the following kernel density estimate:

$$\hat{p}(o_n, x) = \frac{1}{nh^d} \sum_k \sum_j p(o_n, x_j | \mathbf{e}_k, \ell_k) K(\frac{x - x_j}{h}) \quad (7)$$

where the kernel K is a radially symmetric, nonnegative function, centered at zero and integrating to one. From [5], we know that a Mean-Shift search using this formulation will quickly converge to local modes of the underlying distribution. Moreover, the search procedure can be interpreted as a Parzen window probability density estimation for the position of the object center.

From the literature, it is also known that the performance of the Mean-Shift procedure depends critically on a good selection for the kernel bandwidth h. Various approaches have been proposed to estimate the optimal bandwidth directly from the data, e.g. [6, 4]. In our case, however, we have an intuitive interpretation for the bandwidth as a search window for the position of the object center. As the object scale increases, the *relative errors* introduced by equations (4)-(6) cause votes to be spread over a larger area around the hypothesized object center and thus reduce their density in the voting space. As a consequence, the kernel bandwidth should also increase in order to compensate for this effect. We can thus make the bandwidth dependent on the scale coordinate and obtain the following *balloon density estimator* [6]:

$$\hat{p}(o_n, x) = \frac{1}{nh(x)^d} \sum_k \sum_j p(o_n, x_j | \mathbf{e}_k, \ell_k) K(\frac{x - x_j}{h(x)}) \quad (8)$$

For K we use a uniform spherical kernel with a radius corresponding to 5% of the hypothesized object size. Since a certain minimum bandwidth needs to be maintained for small scales, though, we only adapt it for scales greater than 1.0.

We have thus formulated the multi-scale object detection problem as a scale-adaptive Mean-Shift search procedure. Our experimental results in Section 5 will show that this scale adaptation step is indeed needed in order to provide stable results over large scale changes. The performance of the resulting approach depends on the capability of the underlying patch extractor to find image structures that are both typical for the object category and that can be accurately localized in position and scale. As different detectors are optimized for finding different types of structures, the next section evaluates the suitability of various scale-invariant interest point detectors for categorization

4 Influence of Interest Point Detectors

Typically, interest point detectors are only evaluated in terms of their repeatability. Consequently, significant effort has been spent on making the detectors discriminant

Fig. 1. *Scale-invariant interest points found by (from left to right) the exact DoG, the fast DoG, the regular Harris-Laplace, and the fast Harris-Laplace detector on two example images (The smallest scales are omitted in order to reduce clutter).*

enough that they find exactly the same structures again under different viewing conditions. However, we strongly believe that the evaluation should be in the context of a task. In our case, the task is to recognize and localize previously unseen objects of a given category. This means that we cannot assume to find exactly the same structures again; instead the system needs to generalize and find structures that are similar enough to known object parts while still allowing enough flexibility to cope with variations. Also, because of the large intra-class variability, more potential matching candidates are needed to compensate for inevitable mismatches. Last but not least, the interest points should provide a sufficient cover of the object, so that it can be recognized even if some important parts are occluded. Altogether, this imposes a rather different set of constraints on the interest point detector. As a first step we therefore have to compare the performance of different interest point operators for the categorization task.

In this work, we evaluate two different types of scale-invariant interest point operators: the Harris-Laplace detector [15], and the DoG (Difference of Gaussian) detector [14]. Both operators have been shown to yield high repeatability [16], but they differ in the type of structures they respond to. The Harris-Laplace prefers corner-like structures by searching for multi-scale Harris points that are simultaneously extrema of a scale-space Laplacian, while the DoG detector selects blob-like structures by searching for scale-space maxima of a Difference-of-Gaussian. For both detectors, we additionally examine two variants: a regular and a speed-optimized implementation (operating on a Gaussian pyramid). Figure 1 shows the kind of structures that are captured by the different detectors. As can already be observed from these examples, all detectors manage to capture some characteristic object parts, such as the car's wheels, but the range of scales and the distribution of points over the object varies considerably between them.

In order to obtain a more quantitative assessment of their capabilities, we compare the different interest point operators on a car detection task using our extended approach. As a test set, we use the UIUC database [1], which consists of 170 images containing a total of 200 sideviews of cars. For all experiments reported below, training is done on a set of 50 hand-segmented images (mirrored to represent both car directions). In a first stage, we compare the recognition performance if the test images are of the same size as the training images. Since our detectors are learned at a higher resolution than the cars in the test set, we rescale all test images by the same factor prior to recognition. (Note that this step does not increase the images' information content.)

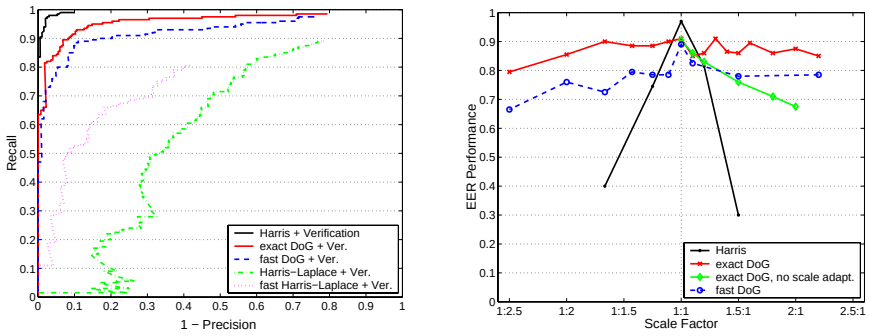

Fig. 2. *Performance comparison on the UIUC database. (left) Precision-Recall curves of different interest point detectors for the single-scale case. (right) EER performance over scale changes relative to the size of the training examples.*

Figure 2(left) shows a comparison of the detectors' performances. It can be seen that the single-scale Harris codebook from [11] achieves the best results with 97.5% equal error rate (EER). Compared to its performance, all scale-invariant detectors result in codebooks that are less discriminant. This could be expected, since invariance always comes at the price of reduced discriminance. However, the exact DoG detector reaches an EER performance of 91%, which still compares favorably to state-of-the-art methods (see Tab. 1). The fast DoG detector performs only slightly worse with 89% EER. In contrast, both Harris-Laplace variants are notably inferior with 59.5% for the regular and 70% for the speed-optimized version.

The main reason for the poorer performance of the Harris-Laplace detectors is that they return a smaller absolute number of interest points on the object, so that a sufficient cover is not always guaranteed. Although previous studies have shown that the Harris-Laplace points are more discriminant individually [7], their smaller number is a strong disadvantage. The DoG detectors, on the other hand, both find enough points on the objects and are discriminant enough to allow reliable matches to the codebook. They are thus better suited for our categorization task. For this reason, we only consider DoG detectors in the following experiments.

5 Robustness to Scale Changes

We now analyze the robustness to scale changes. In particular, we are interested in the limit to the detectors' performance when the scale of the test images is altered by a large factor and the fraction of familiar image structures is thus decreased. Rather than to test individual thresholds, we therefore compare the maximally achievable performance by looking at how the equal error rates are affected by scale changes.

In the following experiment, the UIUC database images are rescaled to different sizes and the performance is measured as a function of the scaling factor relative to the size of the training examples. Figure 2(right) shows the EER performances that can be achieved for scale changes between factor 0.4 (corresponding to a scale reduction of 1:2.5) and factor 2.2. When the training and test images are approximately of the same size, the single-scale Harris codebook is highly discriminant and provides the superior performance described in the previous section. However, the evaluation shows that it is only robust to scale changes up to about 20%, after which its performance quickly drops. The exact-DoG codebook, on the other hand, is not as discriminative and only

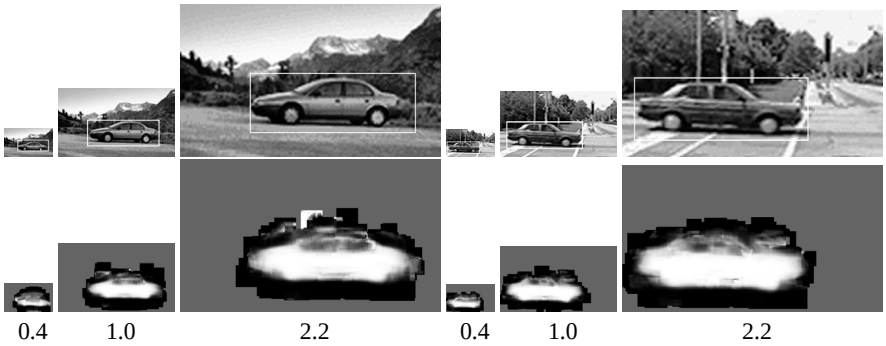

Fig. 3. *(top) Visualization of the range of scales tested in the experiments, and the corresponding car detections. Training has been performed at scale 1.0.; (bottom) Segmentations automatically obtained for these examples. (white: figure, black: ground, gray: not sampled)*

achieves an EER of 91% for test images of the same scale. However, it is far more robust to scale changes and can compensate for both enlargements and size reductions of more than a factor of 2. Up to a scale factor of 0.6, its performance stays above 89%. Even when the target object is only half the size of those seen during training, it still provides an EER of 85%. For the larger scales, the performance gradation is similar. The fast DoG detector performs about 10% worse, mainly because its implementation with a Gaussian pyramid restricts the number and precision of points found at higher scales. Figure 2(right) also shows that the system's performance quickly degrades without the scale adaptation step from Section 3, confirming that this step is indeed important.

An artifact of the interest point detectors can be observed when looking at the performance gradation over scale. Our implementation of the exact DoG detector estimates characteristic scale by computing three discrete levels per scale octave [14] and interpolates between them using a second-order polynomial. Correspondingly, recognition performance is highest at scale levels where structure sizes can be exactly computed (namely $\{0.6, 1.0, 1.3, 1.6, 2.0\}$, which correspond to powers of $(\sqrt[3]{2})$). In-between those levels, the performance slightly dips. Although this effect can easily be alleviated by using more levels per scale octave, it shows the importance of this design decision.

Figure 3 shows a visualization of the range of scales tested in this experiment. Our approach's capability to provide robust performance over this large range of image variations marks a significant improvement over [11]. In the bottom part of the figure, the automatically generated segmentations are displayed for the different scales. Compared to the single-scale segmentations from [11], the segmentation quality is only slightly inferior, while being stable over a wide range of scales.

6 Conclusion and Future Work

In this paper, we have presented a scale invariant extension of the approach from [12, 11] that makes the method applicable in practice. By reformulating the multi-scale object detection problem in a Mean-Shift framework, we have obtained a theoretically founded interpretation of the hypothesis search procedure which allows to use a principled scale adaptation mechanism. Our quantitative evaluation over a large range of scales shows that the resulting method is robust to scale changes of more than a factor of 2. In addition, the method retains the capability to provide an automatically derived object segmentation as part of the recognition process.

As part of our study, we have also evaluated the suitability of different scale-invariant interest point detectors for the categorization task. One interesting result is that, while found to be more discriminant in previous studies [15, 7], the Harris-Laplacian detector on its own does not detect enough points on the object to enable reliable recognition. The DoG detector, on the other hand, both finds enough points on the object and is discriminant enough to yield good recognition performance. This emphasizes the different characteristics the object categorization task brings with it, compared to the identification of known objects, and the consequent need to reevaluate design decisions. An obvious extension would be to combine both Harris-type and DoG-type interest points in a common system. Since both detectors respond to different image structures, they can complement each other and compensate for missing detections. Consequently, we expect such a combination to be more robust than the individual detectors.

Acknowledgments: This work is part of the CogVis project, funded by the Comission of the EU (IST-2000-29375) and the Swiss Federal Office for Education and Science (BBW 00.0617).

References

1. S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV'02*, 2002.
2. D.H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, 1981.
3. E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV'02*, 2002.
4. R. Collins. Mean-shift blob tracking through scale space. In *CVPR'03*, 2003.
5. D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Trans. PAMI*, 24(5):603–619, 2002.
6. D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *ICCV'01*, 2001.
7. G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *ICCV'03*, 2003.
8. R. Fergus, A. Zisserman, and P. Perona. Object class recognition by unsupervised scale-invariant learning. In *CVPR'03*, 2003.
9. A. Garg, S. Agarwal, and T. Huang. Fusion of global and local information for object detection. In *ICPR'02*, 2002.
10. T. Kadir and M. Brady. Scale, saliency, and image description. *IJCV*, 45(2):83–105, 2001.
11. B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV'04 Workshop on Stat. Learn. in Comp. Vis.*, 2004.
12. B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC'03*, 2003.
13. T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.
14. D. Lowe. Object recognition from local scale invariant features. In *ICCV'99*, 1999.
15. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV'01*, pages 525–531, 2001.
16. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *CVPR'03*, 2003.
17. C. Papageorgiou and T. Poggio. A trainable system for object detection. *IJCV*, 38(1), 2000.
18. H. Schneiderman and T. Kanade. A statistical method of 3d object detection applied to faces and cars. In *CVPR'00*, 2000.
19. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR'01*, pages 511–518, 2001.
20. M. Weber, M. Welling, and P. Perona. Unsupervised learning of object models for recognition. In *ECCV'00*, 2000.