Towards Unsupervised Discovery of Visual Categories

Mario Fritz and Bernt Schiele

Multimodal Interactive Systems, TU-Darmstadt, Germany,
{fritz,schiele}@mis.tu-darmstadt.de

Abstract. Recently, many approaches have been proposed for visual object category detection. They vary greatly in terms of how much supervision is needed. High performance object detection methods tend to be trained in a supervised manner from relatively clean data. In order to deal with a large number of object classes and large amounts of training data, there is a clear desire to use as little supervision as possible. This paper proposes a new approach for unsupervised learning of visual categories based on a scheme to detect reoccurring structure in sets of images. The approach finds the locations as well as the scales of such reoccurring structures in an unsupervised manner. In the experiments those reoccurring structures correspond to object categories which can be used to directly learn object category models. Experimental results show the effectiveness of the new approach and compare the performance to previous fully-supervised methods.

1 Introduction

Over the years various approaches have been proposed for the recognition of object categories often based on models learned directly from image data. The approaches, however, vary greatly in the specific task they address: from simple present/absent decision [1, 2] over object class detection and localization [3] to pixel-level segmentation [4]. In this paper we deal with the problem of object detection and localization. Another difference between the proposed methods is the amount of supervision used and provided for the training data. The types of annotation varies from pixel-level segmentations [5], over bounding-box annotations [3] to unsupervised methods [2, 6, 7]. Very recent approaches for learning multiple categories do not even require the information which category is presented in which image [8]. While approaches using more supervision tend to require less training data, there is a clear desire to use less supervision typically at the price to use more unlabeled training data.

The central problem addressed in this paper is to discover and learn objects category models as reoccurring patterns of local appearance in sets of training data. It may seem quite unrealistic to discover object categories in this way. However, many appearance-based approaches explicitly or implicitly rely on the fact that both the local appearance as well as its structural layout exhibit reoccurring patterns that can be learned and modeled (e.g. [2, 4, 9]). A key idea

of our approach is therefore to discover reoccurring patterns in multiple images without the model of any particular object. Finding the locations and scales of such reoccurring structures effectively corresponds to unsupervised annotations of the training data. As we will show, the proposed approach enables effective object class discovery in unlabeled images. Using those estimated annotations a model of an object class can be learned.

Learning object models in an unsupervised fashion may be formulated in one single EM-loop as in e.g. Fergus et al [2]. In that method, appearance and structure are learned simultaneously making the learning computationally expensive and thus restricting the complexity of the model. Recently a new approach for object discovery has been proposed based on a pLSA-model [8]. Since the underlying model is a bag-of-word representation, the object discovery is based on local appearance alone neglecting structural information. [7] extends the initial approach to also include some structural information on top of the pLSA model, but the object discovery is still based on appearance only.

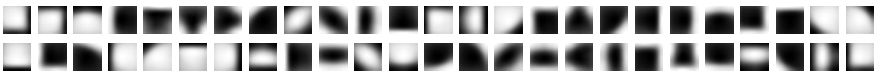
The main contributions of this paper are the following: First, we propose a novel scheme to discover object category members in images, which is based on the idea of estimating the locations and scales of reoccurring patterns. The estimates can be seen as an automatic annotation procedure of the training data. Second, we experimentally show the applicability of this idea for object discovery on several object classes. Third, we use the estimated annotations to learn object class models for object detection and localization. Fourth, we analyze the performance of such object class models on standard datasets.

The paper is organized as follows: Section 2 describes a method for locating reoccurring structure for which in Section 3 we present a method to robustly estimate the intrinsic scale of the associated objects. Section 4 shows how a model like [4] can be learnt from the estimated annotations. Finally, we show in Section 5 the usefulness of the obtained information on a image ranking and an object detection task.

2 Object Discovery

Our new approach to unsupervised object discovery is based on efficiently finding reoccurring spatial patterns of local appearances in a set of training images. We use a generic codebook representation which is also the basis of the object scale estimation procedure as presented in Section 3.

Generic Codebook Representation Similar to other approaches for recognition [1], material classification [10] and detection [4], we use an initial clustering procedure to obtain a visual codebook. Since we do not want to assume a priori information on parts or common structure of the object category, we use a fixed generic codebook produced on unrelated background images. We extract image patches on the Caltech background images [11] using a scale-invariant Hessian-Laplace interest point detector [12]. Those image patches are clustered by k-means using normalized gray-scale correlation as similarity measure. The result looks as follows:



Scale-Invariant Patterns We define a pattern $\Psi_{k,r}$ in image k with reference point r to be characterized by a set of distributions $\{p(h|\Psi_{k,r,c})\}_c$. Each of the $p(h|\Psi_{k,r,c})$ encodes the spatial distribution of the features in image k that match to a certain codebook c. The coordinates $h = (h_x, h_y)$ are scale-normalized with the intrinsic feature scales σ (obtained from the scale-invariant interest point detector) and computed relative to a reference point $r = (r_x, r_y)$

$$h = \left(\frac{x - r_x}{\sigma} \, , \, \frac{y - r_y}{\sigma} \right). \tag{1}$$

Using this scale-normalized coordinates is beneficial, as the pattern becomes characteristic for the particular reference point r. This allows to locate reoccurring patterns even though they appear at different global scales.

Method We formulate the unsupervised discovery of reoccurring spatial patterns of local appearances as finding for each image the most likely pattern given all observed patterns in the training data. Therefore we are interested in finding the reference point $\hat{q}_j$ associated with the most likely pattern in each image j given all observed patterns $\mathbf{\Psi} = \{\Psi_{k,r}\}_{k,r}$

$$\hat{q}_j = \arg\max_q p(\Psi_{j,q}|\mathbf{\Psi}). \tag{2}$$

To simplify notation, the reference points q and r are assumed to be quantized. The likelihood estimate is obtained by marginalizing over the codebook entries c, scale-normalized coordinates h, reference points r, and images k

$$p(\Psi_{j,q}|\mathbf{\Psi}) = \sum_c \sum_h \sum_r \sum_k p(\Psi_{j,q,c}|h) p(h|\Psi_{k,r,c}) p(\Psi_{k,r,c}).$$

Using Bayes' formula we obtain

$$p(\Psi_{j,q,c}|h) = \frac{p(h|\Psi_{j,q,c}) p(\Psi_{j,q,c})}{p(h)}. \tag{3}$$

By assuming uniform priors, $p(\Psi_{k,r,c})$ and $p(h)$ can be written as constant $\frac{1}{Z}$. This assumption is justified, by a uniform partitioning of our data using k-means clustering. Eq. 3 simplifies to

$$p(\Psi_{j,q}|\mathbf{\Psi}) = \frac{1}{Z} \sum_c \sum_h \sum_r \sum_k p(h|\Psi_{j,q,c}) p(h|\Psi_{k,r,c}). \tag{4}$$

An example of this likelihood estimate on the multi-scale TUD motorbikes [11] is overlaid on one of the images in Figure 1 as iso-lines. In this image we can clearly see two maxima which correspond to two motorbikes.

Eq. 4 can be interpreted as collecting evidence for pattern $\Psi_{j,q}$ with respect to all other patterns $\mathbf{\Psi}$ by searching for matching feature with appearance c

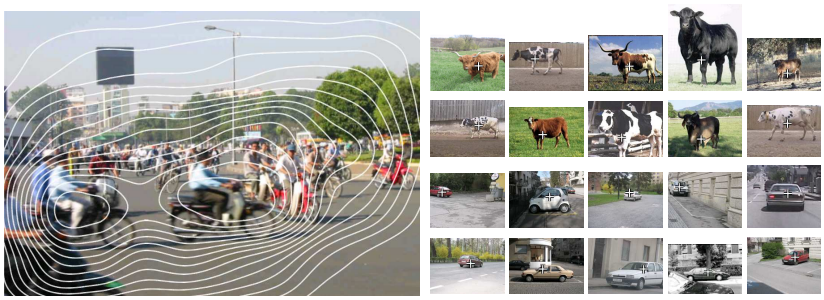


Fig. 1. *(Left) Example of the computed likelihood on the multi-scale TUD motorbikes. (Right) Example result of our procedure for object discovery on car and cow images including varying position, scale and viewpoint and heterogeneous background.*

and scale-normalized position h. Although this seems computationally infeasible, we introduce an efficient method to evaluate eq. 4 using scale-invariant feature hashing - similar to the idea of geometric hashing [13]. The idea is to index all features of the image database by quantized scale-normalized coordinates h, and store them in the hashes $\mathcal{H}_c$ for each matching codebook cluster c. Features which are similar in appearance and scale-normalized position h are now stored in the same hash bin. More importantly, the matches can be used to backproject the support of all patterns $\Psi_{j,q}$ with respect to all patterns. As a result, all $p(\Psi_{j,q}|\mathbf{\Psi})$ given by the complex eq. 4 can be computed by a single loop over the hash bins of hashes $\mathcal{H}_c$.

Evaluation To test the proposed procedure for object discovery with respect to robustness against translation, scaling, occlusion, and background clutter we ran tests on three object categories: motorbikes, cows, and cars. For the cows we used the training set of the TUD cows [11], as well as the cows from [14]. For the cars we used the training set of the PASCAL challenge [11]. Examples for the estimated object centers are shown in Figure 1. Despite the strong variations in appearance and view-point, the objects were successfully localized. The reference locations were quantized on a 10×10 grid.

To gain more insights, we perform a more detailed quantitative analysis on the Caltech motorbike training set [11] which consists of 400 images. We compute the distance between our estimate and the center of the groundtruth bounding box annotation normalized by the object width. The average distance is 0.10, which we consider to be very good, as the groundtruth annotations are not really accurate themselves. Nearly all errors are below a normalized distance of 0.3, which is well below the noise level assumed in the evaluation of the scale estimation method in Section 3.

3 Object Scale Estimation

From the procedure for object discovery described in the previous section we obtain localized patterns $\Psi_{j,q}$ at reference points $\hat{q}_j$ for each image j. However,

since these reoccurring patterns are obtained in a scale-invariant fashion, they are of unknown scale s. While it is advantageous, that no explicit knowledge of the object scale is required for discovering reoccurring patterns, tasks like training an object model for detection need an estimate of the object scale to learn a model across the training instances.

Method The proposed method matches scale-invariant patterns to collect evidence for their associated global object scale. Different methods to obtain a robust estimate are proposed and evaluated. As the absolute, global object scale only exists with respect to a reference scale, we formulate the scale estimation problem as finding the pairwise relative scale $\hat{\rho}_{k,l} = s_k/s_l$ between two discovered patterns Ψ_k and Ψ_l in a pair of images k and l. In analogy to eq. 2 we describe the problem of finding the most likely relative scale $\hat{\rho}_{k,l}$ with respect to the two patterns of the image pair as

$$\hat{\rho}_{k,l} = \arg\max_{\rho_{k,l}} p(\rho_{k,l}|\Psi_k, \Psi_l) \tag{5}$$

We assume that for matching features the ratio of the intrinsic scale σ of the matched structures is equal to the ratio of the global scales s between the patterns and their associated objects $\rho_{k,l} = s_k/s_l = \sigma_k/\sigma_l$. According to this we factor eq. 5 and marginalize over the codebook entries c and the scale-normalized coordinates h

$$p(\rho_{k,l}|\Psi_k, \Psi_l) = \sum_{\sigma_l} p((\rho_{k,l}\sigma_l)|\Psi_k)p(\sigma_l|\Psi_l) = \sum_c \sum_h \sum_{\sigma_l} p((\rho_{k,l}\sigma_l), h|\Psi_{k,c})p(\sigma_l, h|\Psi_{l,c})$$

As in Section 2 we store all features in the hashes $\mathcal{H}_c$. Our efficient data structure allows to compute all these likelihoods in one loop over the hash bins.

The estimates from eq. 5 can be interpreted as a fully connected graph, where the patterns in the images are the nodes and the relative scales of the patterns are attached to the edges. To make our method robust with respect to outliers, we compute confidence scores for all estimated relative scales. These are computed by indentifying image triplets with consistent relative scale estimates: Given three images I_a, I_b, I_c with their relative scales $\rho_{a,b}, \rho_{b,c}, \rho_{a,c}$, the confidence for all three scale estimates is increased if the equation $\rho_{a,b}\rho_{b,c} = \rho_{a,c}$ is fulfilled.

In this paper we investigate three different methods to derive a unique scale estimate for each pattern from the pairwise relative scale information: *least squares*, *maximum spanning tree*, and *min-linkage method*.

The *least squares method* is based on a linear system of equations to estimate the unknown scales without using the computed confidences. Considering two patterns Ψ_k, Ψ_l with the global scale of the patterns s_k, s_l of the associated object instances, we compute a least-squares fit for the global scales s from all the estimated relative scale according to:

$$\frac{s_k}{s_l} = \rho_{k,l} \implies \log s_k - \log s_l = \log \rho_{k,l}. \tag{6}$$

This method is comutational expensive, because the number of equations grows quadratically in the number of images, and its estimes are sensitive to outliers.

The *maximum spanning tree method* computes a maximum spanning tree on the graph of confidences. The scale estimates can be directly computed from this tree by fixing one scale. Although this method has low computational complexity, the estimates are rather unstable, as shown in Section 3.

As a compromise between efficient computation and robust estimation, we propose a third method. The *min-linkage method* considers for every image the n most confident relative scales to all other images and therefore the number of equations grows only linearly with the number of images. The estimate of the scales is still robust due to the least-squares estimation.

The above described methods estimate relative scales, however, for the detection experiments (Section 5) an absolute scale based on the extent of the object is required. One possibility is to specify a reference scale for one image. In the experimental evaluation it turned out that this is not necessary, as the absolute object radius can be chosen to be twice the mean feature distance to the center *after* aligning all objects with the computed relative scale estimates.

Evaluation To evaluate the accuracy of our new scale estimation scheme, we again use the Caltech motorbike database with annotated object centers, which has a scale variation of 2 octaves. The mean deviation of the estimated scales from the true scales is roughly $\frac{1}{9}$ of an octave for the least-squares and $\frac{1}{5}$ for the min-linkage method (minimum linkage preserves $40 = 10\%$ of the most confident scales). Additionally, we evaluated the robustness of the system with respect to Gaussian noise in the center point annotation. Even when the noise is amplified until the 3σ-radius reaches $\frac{2}{3}$ of the object radius - which is twice the noise level we measured for the center point estimate in Section 2 - the mean deviation of the estimated scales from the true scale is roughly $\frac{1}{4}$ of an octave for least-squares and minimum linkage. The maximum spanning tree method reaches this error already at half the noise level. As a conclusion we use the minimum linkage method in our following experiments, as it shows about the same accuracy as the full least-squares, but with a much lower computational cost.

4 Model Estimation for Detection

Object Category Detection with ISM The *Implicit Shape Model (ISM)* [5] is a versatile framework for scale-invariant detection of object categories, which has shown good performance on challenging detections tasks [11]. It uses a flexible non-parametric representation for modeling visual object categories by spatial feature occurrence distributions with respect to a visual codebook. For details we refer to [5]. Additionally the method allows for back-projecting the support of the hypotheses to infer figure-ground segmentation masks and performing an MDL-based reasoning to resolve multiple and ambiguous hypotheses [4]. However, the generation of an object specific visual codebook and the MDL-based reasoning step require figure-ground segmentations for the training images which introduce high annotation effort.

Unsupervised Learning of Models for Detection One of our contributions is to show that one can achieve high recognition performance by using

the estimated center point (Section 2) and scale (Section 3) instead of manually produced segmentations. As we do not have a detailed segmentation mask at our disposal when using those location and scale estimates, we use a simple but (as will be seen in the experiments) effective approximation. Figure 3 shows our rough approximation by assuming the segmentation to be a circle specified by the estimated center point and scale.

To learn the ISM model, we first switch from the generic codebook (Section 2) to an object specifc SIFT representation [15] computed on Hessian-Laplace interest points [12]. We use the approximated segmentation (circles) to determine the object features for clustering. Given the approximated segmentation and the new codebook, we can proceed training the ISM as described in [5]. Despite the crude approximation of the segmentations with circles, it is possible to infer segmentations for the hypothesis on test images as shown in Figure 3.

5 Experiments

Whereas the previous sections analyzed the proposed object discovery and object scale estimation separately, this section shows the applicability to image ranking and object category detection. While the ranking task also shows the scalability to large numbers of images, the detection experiments evaluate how the proposed method generalizes to different categories. In addition we will show that the approximated segmentation masks from Section 4 are effective and even crucial to obtain high level detection performance.

Image Ranking In the following, experiments we show that the proposed method for unsupervised object discovery from Section 2 can be used on its own for an image ranking task. Using a keyword search for motorbikes we downloaded 5246 images containing a wide range of different motorbike types (e.g. cruiser, sportbike, touring, scooter, moped, off-road, combination) captured from different viewpoints. Naturally quite a number of those images only show close-ups, parts or even unrelated objects. Our task is to sort these images out. We use our method for object discovery to rank the images by the likelihood (eq. 4). Note, that this ranking is obtained in an totally unsupervised and no validation set as in [7] is needed. Figure 4(left) shows the ROC curves obtained by running our approach with and without spatial information. If the spatial information of the features is discarded, our representation reduces to a bag-of-words representation. The use of spatial information improves the results significantly, which demonstrates the improvment of our model over purely appearance-based approaches. Qualitative results for our new approach using appearance and spatial structure are shown in Figure 2. As scooters were the dominating motorbike type in the set (1169 of 5246), they also appear first in the ranking.

Visual Category Detection Task In the detection experiments we train a model according to Section 4 and use it to localize objects in the test images. Detections are only accepted as correct if the hypothesized bounding box fits the groundtruth annotation. Multiple detections are counted as false positives. For better comparability we use the acceptance criterion described in [16]. We want

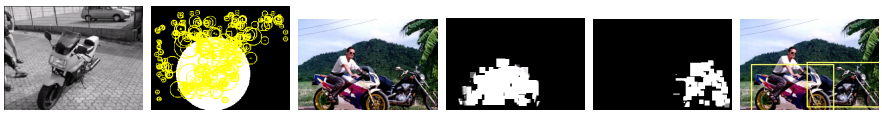
Fig. 2. *The proposed method for Object Discovery also facilitates ranking of the images. (left) best ranked images (right) worst ranked images.*

to emphasize, that no parameters had to be tuned for the proposed approach for unsupervised learning. In terms of efficiency, the approach for object discovery can estimate object locations in 200 images in 11 minutes on a 3Ghz Pentium4, whereas the object scale estimation takes 6 minutes.

Unsupervised Learning for Detection. Figure 4(middle) shows results on the multi-scale TUD motorbike test set [11], which includes significant scale variations, partial occlusions and multiple instances per image. The models are trained on the Caltech motorbikes [11]. The best results published on this data-set are 80% EER using accurate pixel-level segmentation and ISM (supervised training with MDL) and 81% adding an additional SVM-stage (supervised training with MDL+SVM) [16]. Quite importantly, the performance of the proposed unsupervised object discovery method (specific SIFT codebook with MDL - 150) is very similar to the supervised training of ISM. The EER of 81% can be further increased to 84% by using 400 instead of 150 training images (again in an unsupervised fashion) and which is the best performance presented so far on the test set. Compared to the SVM approach [16] the precision is slightly worse, but the achievable recall is higher. So adding an SVM classifier in a similar fashion has the potential to further increase the overall performance. Overall the results are highly encouraging as they indicate that high annotation effort can be replaced by using a larger amount of training data.

Evaluation of the Approximated Segmentation Masks. Figure 3 shows a test image with the estimated segmentation masks and the final detections. While the mask is far from being perfect, the computed support of the hypotheses is approximately correct. Figure 4(middle) shows how the performance increases significantly when this approximation is used to perform the MDL-based hypothesis verification. The results support our claim, that the estimated segmentation masks are accurate enough and facilitate the training of a model that gives competitive performance. The figure also shows the importance of switching to an object class specific SIFT codebook (Section 4).

Generalization to other Categories. To investigate how this approach generalizes to other categories and compare our method to previous work, we conduct experiments on cows, faces, and cars. The results are reported in Figure 4(right). The training sets TUD cows and Caltech faces [11] are selected, as they include



(a) (b) (c) (d) (e) (f)

Fig. 3. *(a) training image (b) estimated approximation of object segmentation (c) test image (d)+(e) inferred segmentation for hypothesis (f) final detections*

a significant amount of variation of the object position in the training data to underline the performance of the proposed method for object discovery. For the cows we use the same test setting as in the supervised approach of [16]. Our unsupervised approach achieves an equal error rate performance of 79.9% whereas the supervised reference achieved 93.2% [16]. As the background is for some training images the same, we learnt it as reoccurring structure. As it is part of the model, we get some strong hypotheses on these background structures which also occur in the test set and that are responsible for the decrease in performance. On the UIUC car and caltech face database we compare to the unsupervised method of Fergus [2] On the cars we get an equal error rate performance of 89.5% in comparison to 88.5% in [2] using the same evaluation criterion. We achieve this performance training on only 50 car images and their mirrored versions from the TUD car database [11]. The best performance on this dataset is reported by the supervised method in [4] achieving 97% equal error rate performance. In [17] a detection performance for the model of [2] of 78% equal error rate is presented on the caltech face database. Our approach achieves a significant improvement by an equal error rate performance of 81.1%.

6 Conclusion

We have proposed an efficient and flexible framework for discovering visual object categories in an unsupervised manner which makes use of appearance and spatial structure at the same time. The approach is based on two new components for object discovery and object scale estimation, that extract information about reoccurring spatial patterns of local appearance. The experimental results show that our system facilitates unsupervised training of an model for object class detection that has equal or even better performance than previous unsupervised approaches. In addition, the method was used to rank images without any supervision or validation. Results are presented on a large image database of over 5000 images including a significant amount of noise. Finally, we obtained comparable results w.r.t. a stronly supervised state-of-the-art detection system on a challenging multi-scale test set. We showed that we can compensate for the decrease in performance by adding more training examples, which results in the best performance shown so far on this test set.

Acknowledgments: This work has been funded, in part, by the EU project CoSy (IST-2002-004250).

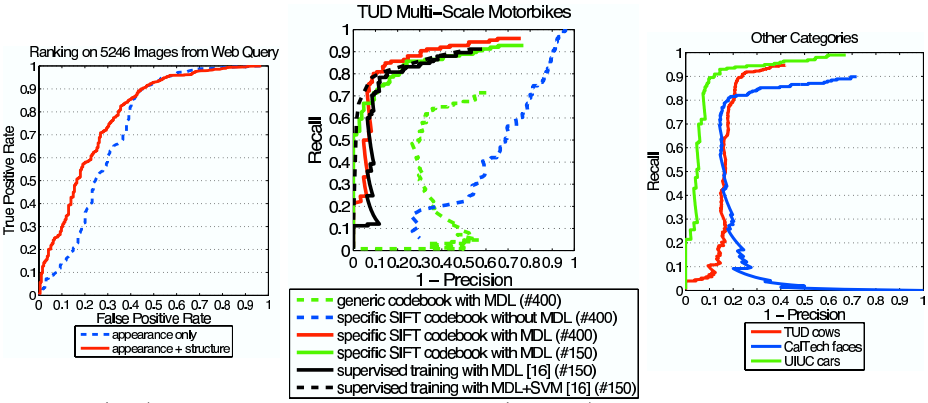


Fig. 4. *(left) ROC-curve of ranking task (middle) performance comparison to supervised baseline (right) generalization to other categories and data sets*

References

1. Csurka, G., Dance, C., Fan, L., Willarnowski, J., Bray, C.: Visual categorization with bags of keypoints. In: ECCV'04 Workshop on Stat. Learn. in Comp. Vis., Prague, Czech Republic (2004) 59–74
2. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR'03, Madison, WI (2003)
3. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR'01. (2001)
4. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV'04 Workshop on Stat. Learn. in Comp. Vis., Prague, Czech Republic (2004) 17–32
5. Leibe, B., Schiele, B.: Scale invariant object categorization using a scale-adaptive mean-shift search. In: DAGM'04, Tuebingen, Germany (2004)
6. Winn, J.M., Jojic, N.: Locus: Learning object classes with unsupervised segmentation. In: ICCV. (2005) 756–763
7. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: ICCV'05, Beijing, China (2005)
8. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their locations in images. In: ICCV'05, Beijing, China (2005)
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. IJCV **61**(1) (2005) 55–79
10. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV **43**(1) (2001) 29–44
11. Pascal: The PASCAL Object Recognition Database Collection (2005) http://www.pascal-network.org/challenges/VOC.
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10) (2005) 1615–1630
13. Wolfson, H.J., Rigoutsos, I.: Geometric hashing: An overview. IEEE Comput. Sci. Eng. **4**(4) (1997) 10–21
14. Hillel, A.B., Hertz, T., Weinshall, D.: Efficient learning of relational object class models. In: ICCV'05, Beijing, China (2005)
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004) 91–110
16. Fritz, M., Leibe, B., Caputo, B., Schiele, B.: Integrating representative and discriminant models for object category detection. In: ICCV'05, Beijing, China (2005)
17. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. (2005) Under review