Object Localization with Boosting and Weak Supervision for Generic Object Recognition

Andreas Opelt and Axel Pinz

Institute of Electrical Measurement and Measurement Signal Processing,
Graz University of Technology, Austria
{opelt, pinz}@emt.tugraz.at

Abstract. This paper deals, for the first time, with an analysis of localization capabilities of weakly supervised categorization systems. Most existing categorization approaches have been tested on databases, which (a) either show the object(s) of interest in a very prominent way so that their localization can hardly be judged from these experiments, or (b) at least the learning procedure was done with supervision, which forces the system to learn only object relevant data. These approaches cannot be directly compared to a nearly unsupervised method. The main contribution of our paper thus is twofold: First, we have set up a new database which is sufficiently complex, balanced with respect to background, and includes localization ground truth. Second, we show, how our successful approach for generic object recognition [14] can be extended to perform localization, too. To analyze its localization potential, we develop localization measures which focus on approaches based on Boosting [5]. Our experiments show that localization depends on the object category, as well as on the type of the local descriptor.

1 Introduction

There is recent success in weakly supervised object categorization from input images (e.g. [4], [14], [8]). Systems can learn based on given piles of images containing objects of certain categories, and piles of counterexamples, not containing these objects. These approaches cope well with the generalization over an object category and perform well in categorization. There are two main aspects in analyzing these approaches with respect to object localization. First, the data needs to be complex enough to challenge a system regarding its localization performance. Second, it is important to discuss the amount of used supervision. Clearly the task of localization becomes easier when one uses a high degree of supervision (e.g. the segmented object) to train the classifier. One might argue that a high degree of supervision during training is similar to human categorization behavior, as humans can easily separate the object of interest from the background. We are interested in designing a vision system that can learn to localize categories with the lowest possible amount of supervision, which should be useful for a broad variaty of applications.

H. Kalviainen et al. (Eds.): SCIA 2005, LNCS 3540, pp. 862–871, 2005.

In [14] we presented an approach which uses almost no supervision (just the image labels) and also performs well on complex data. This combination brings up the question of localization. As we use Boosting [5] in our categorization approach we want to focus on measuring localization perfomance related to that learning technique. There are two main contributions of this paper: First, we set up a new image database which is sufficiently complex, balanced, and provides localization ground truth[1]. Second, we define and incorporate localization measures that correspond with the feature selection process during the learning step (based on AdaBoost [5]). object.

2 Related Work

The extensive body of literature on generic object recognition reduces if one is also interested in localization. The first group of approaches deals with a tradeoff between generic classification with low supervision and localization performance with higher supervision (e.g. [2], [4], [16]) generally on easier data. Other approaches are really good in localization but just for specific objects (e.g. [15], [8]). Subsequently, we discuss some of the most relevant and most recent results with special emphasis of the problem of localization. The method introduced by Lazebnik et al. [8] is based on semi-local affine parts which are extracted as local affine regions that stay approximately affinely rigid over several different images of the object. The localization performance of that approach is good, but in contrast to our approach they focus on specific object recognition.

In [4] Fergus et al. presented their recent success on object categorization using a model of constellations of parts learned by an EM-type learning algorithm. This leads to a very good recognition performance, but their training images do not have the complexity to show the difficulties in localization with weak supervision. Compared to that our data is highly complex. Learning object relevant data with low supervision from highly cluttered images was discussed by Ruthishauser et al. [15]. On our data their attention algorithm did not work so well. Also the authors do specific object recognition whereas we try to solve the generic problem.

The work by Agarwal et al. [1] solves the problem of localization in a very elegant manner. They localize cars viewed from the side by detecting instances of a sparse, part-based representation. However, they learn their model from sample image portions, which are cut out and show just the objects themselves. In this sense, their approach should be regarded as highly supervised with respect to localization.

Leibe and Schiele [10] also use a sparse, part-based representation forming a codebook for each category. But they add an implicit shape model which enables them to automatically segment the object as a result of their categorization. Having these segments means also that the object is localized. This approach is also scale invariant. In a similar manner as for [1], we notice that localization is less difficult due to the higher degree of supervision in using easier training images.

[1] The database used in [14] was not balanced concerning the background in the positive and negative images.

3 Method and Data

3.1 Database and Localization Ground Truth

We have set up a new image database ("GRAZ-02"[2]). It contains four categories: Person (P), Bike (B) and Cars (C) and counterexamples (N, meaning that it contains no bikes, no persons and no cars). Figure 1 shows two example images for each of the four categories. This database is sufficiently complex in terms of intra class variation, varying illumination, object scale, pose, occlusion, and clutter, to present a challenge to any categorization system. It is also balanced with respect to background, so that we can expect that a significant amount of learned local descriptors should be located on the objects of interest. So the backdoor of categorizing images of e.g. cars by searching for traffic signs and streets is not easily possible. All relevant objects in all images of categories P, B, and C have been manually segmented. This is not used for training, but provides a localization ground truth which is required for experimental evaluation. Some examples are shown in figure 2.

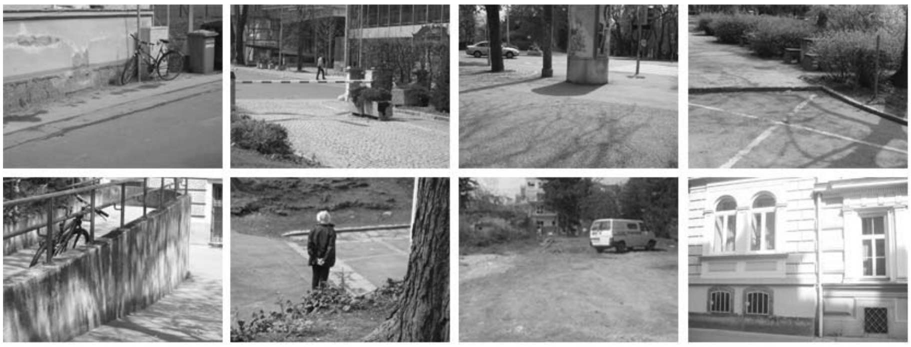

Fig. 1. Two example images for each category of our database (all of these images were correctly categorized using our approach from [14]). Column 1: Bikes (B), 2: Persons (P), 3: Cars (C), 4: counter-class (N)

3.2 Image Categorization

We build on our categorization framework first introduced in [14]. Its localization abilities should be studied, because it shows good results on complex images with no other supervision than the image labels. It is briefly summarized here as a prerequisite to understand the subsequent sections on localization and learning. To train a classifier, the learning algorithm is provided with a set of labeled training images. Note that this is the only amount of supervision required. The object(s) are not pre-segmented, and their location and pose in the images are unknown. The output of the learning algorithm is a

[2] The database and the ground truth are available for download at: http://www.emt.tugraz.at/∼pinz/data

final classifier $H(I) = sign(\sum_{j=1}^{T} h_j(I)w_{h_j})$ (further on also called "final hypothesis") which predicts if a relevant object is present in a new image I. It is formed by a linear combination of T weak classifiers h_j each weighted by w_{h_j}. The output of a weak classifier on an image I is defined as: $h_j(I) = 1$ if $d_M(h_j(I), p_I) \leq th_{h_j}$ and $h_j(I) = 0$, otherwise. Here th_{h_j} denotes the classifiers threshold and $d_M(h_j(I), p_I)$ defines the minimum distance (here we use the Euclidean distance for SIFTs and Mahalanobis distance for the other descriptors) of the weak classifier $h_j(I)$ (also called "weak hypothesis") to all patches p_I in an image I. For details on the algorithm see [14] and [5]. The learning procedure works as follows: The labeled images are put into a preprocessing step that transforms them to greyscale. Then two kinds of regions are detected. Regions of discontinuity are the elliptic regions around salient points, extracted with various detectors (Harris-Laplace [12], affine Harris-Laplace [13], DoG [11]). Regions of homogeneity are obtained by using Similarity-Measure-Segmentation [6], and Mean-Shift segmentation [3]. Next, the system calculates a number of local descriptors of these regions of discontinuity and homogeneity (basic moments, moment invariants [7], SIFT descriptors [11], and certain textural moments [6]). These detection and description methods can be combined in various ways. AdaBoost [5] is used as learning technique. The result of the training procedure is saved as the final hypothesis. A new test image I is categorized by calculating the weighted sum $H(I)$ of the weak hypotheses that fired. Firing means that $d_M(h_j(I), p_I) < th_{h_j}$, as mentioned before. An overview of the image categorization system is depicted inside the framed part of figure 3.

3.3 Object Localization

So far the system is able to learn and to categorize. Now, we extend it aiming at object localization and at the possibility to measure the localization performance. Figure 3 shows the extended framework with all additional components highlighted in grey.

Image categorization is based on a vector of local descriptors (of various types, see section 3.2). They can be located anywhere (around salient points or homogeneous regions) in the image. These categorization results lack a systematic investigation in terms of object localization. Which patches are located on the objects, which ones on the background? How is the relation of object vs. background patches? To answer these questions, we define two localization measures λ_h and λ_d, which correspond with the way, features are selected and weighted by AdaBoost.

λ_h evaluates the localization abilities of a *learned final hypothesis*:

$$\lambda_h = \frac{\sum_{j=1}^{T}(w_{h_j} | d_M(h_j, p_I) < th_{h_j}, p_M \in obj)}{\sum_{j=1}^{T}(w_{h_j} | d_M(h_j, p_I) < th_{h_j}, p_M \notin obj)} \tag{1}$$

Where p_M is defined as the patch in an image I with the minimum distance to a certain hypothesis h_j, and "obj" is the set of points forming the ground truth of image I (i.e. the pixel coordinates of the segmented object in the ground truth data). Thus, a large value of λ_h depicts a situation, where many patches of a final hypothesis are located on the object, and few ones in the background.

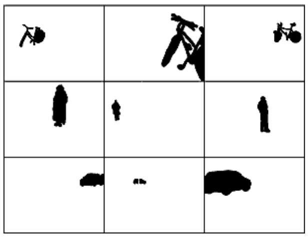

Fig. 2. Ground truth examples. Row 1: Bikes, row 2: Persons, row 3: Cars

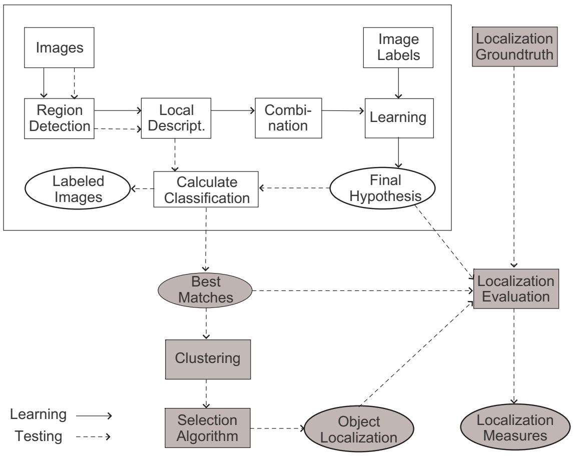

Fig. 3. Our original categorization framework ([14], shown inside the frame), and the extensions for object localization (highlighted in grey)

λ_d evaluates the localization in a *test case*:

$$\lambda_d = \frac{\sum_{i=1}^{m} c(I_i|obj)}{\sum_{i=1}^{m} c(I_i|bg)} \tag{2}$$

with

$$c(I_i|X) = \begin{cases} 1 \text{ if } \sum_{j=1}^{T}(w_{h_j}|d_M(h_j,p_I) < th_{h_j}, p_M \in X) \\ \quad > \sum_{j=1}^{T}(w_{h_j}|d_M(h_j,p_I) < th_{h_j}, p_M \notin X). \\ 0 \text{ else} \end{cases}$$

Where m is the number of test images I and X is a set of pixels obtained from ground truth data (we again use "obj" for the set of pixels belonging to the object (ground truth), and "bg" for the others). Thus, λ_d calculates the ratio of the images categorized mainly by object relevant data versus the number of images categorized mainly by contextual information.

λ_h enables us to estimate the learned localization abilities, and λ_d gives us an accumulated object localization quality for a number of test cases. But we are also interested in individual localization results. To obtain the localization of the object in a specific test image I we compute the positions of the best matching description vectors for each weak hypothesis, and calculate spatial clusters using kmeans [3] (see figure 3). Having k clusters $C_{cl}, cl = 1, \ldots, k$, the difficult task is to find out which one represents the object location. Our straightforward 'Selection Algorithm' consists of the following steps:

[3] One could also use agglomerate clustering here. This would avoid setting a fixed parameter k, but would introduce the need of a threshold for the agglomerations. However, we set k to relative small numbers and got good results.

1. Calculate cluster weights $W_{cl} = \sum_{j=1}^{T}(w_{h_j}|d_M(h_j, p_I) < th_{h_j}, p_M \in C_{cl})$ for every cluster $cl = 1, \ldots, k$.
2. Count the number of best matches P_{cl} in each cluster.
3. Set a cluster rectangle R_{cl} covering all cluster points for each cluster.
4. Increase the rectangle size by e pixels on each side.
5. Select the cluster C_{max} where both, W_{cl} and P_{cl} have the highest value. If no such cluster is available take the one where P_{cl} is maximal (we found that using P_{cl} instead of W_{cl} gives better results).
6. If $R_{C_{max}}$ intersects with other R_{cl} extend $R_{C_{max}}$ to cover the intersecting R_{cl}.
7. If $R_{C_{max}}$ is closer than d pixels to another cluster R_{cl} extend $R_{C_{max}}$ to cover the intersecting R_{cl}.
8. Go back to 6. and iterate l times. If either l is reached or no further changes occured in steps 6. and 7. exit with $R_{C_{max}}$ as object location.

This algorithm delivers an object location in a test image I which is described by the coordinates of a rectangle $R^I_{C_{max}}$. Note that multiple object detection in one image is not possible without a spatial object model. If our data contains multiple objects (just some cases) we aim for the detection of one of the object instances. To measure this effective localization performance we use the evaluation criterion proposed by Agarwal et al. [1]. It describes that the object has to be located within an ellipse which is centered at the true location. If (i', j') denotes the center of the rectangle corresponding to the true location (ground truth) and (i, j) denotes the center of our rectangle $R_{C_{max}}$ then for (i, j) to be evaluated as correct detection it requires to satisfy

$$\frac{(i-i')^2}{\alpha_{height}^2} + \frac{(j-j')^2}{\alpha_{width}^2} \leq 1, \tag{3}$$

where $\alpha_{height}, \alpha_{width}$ denote the size of the ellipse. Note that we do not use the measure for a multiscale case as Agarwal et al., because we need to cope with training objects at varying scales.

4 Experiments and Results

4.1 Parameter Settings

The results were obtained using the same set of parameters for each experiment. All the parameter settings regarding the learning procedure are similar to the ones we used in [14] and [6]. The tresholds for reducing the number of salient points are set to $t_1 = 30000$ and $t_2 = 15000$.

For the localization method we used $k = 3$ cluster centers. For the selection algorithm the following parameters were used: $e = 20$, $d = 10$ and $l = 2$. For the evaluation criterion of Agarwal et al. [1] we used $\alpha_{height} = 0.5 \cdot h_{R_{GT}}$ and $\alpha_{width} = 0.5 \cdot w_{R_{GT}}$ with $h_{R_{GT}}$ and $w_{R_{GT}}$ being the height and width of the box delimiting the ground truth of an image.

4.2 Image Categorization

For comparison with other approaches regarding categorization, we used the Caltech database. We got better or almost equal results on this rather easy dataset (classification rates ranging beween 90% and 99.9%, for details see [14], [6]). From our database we took a training set consisting of 150 images of the object category as positive images and 150 of the counter-class as negative images. The tests were carried out on 300 images half belonging to the category and half not[4]. Table 2 shows the categorization results measured in ROC-equal-error rates of various specific combinations of region extractions and description methods on the three categories of this database. The average ratio of the size of the object versus the image size (counted in number of pixels) is: 0.22 for Bikes, 0.17 for Persons and 0.09 for Cars.

4.3 Localization and Localization Measures

Localization performance on easy datasets is good. For example on motorbikes (Caltech) localization gets results above 90%. This data shows the object highly prominent with just little background clutter, what reduces the localization complexity. We thus proceed by presenting localization results for our more complex GRAZ-02 dataset. The left half of table 1 shows the values of the measure λ_h for the various techniques (the same as in table 2) on all three categories. Comparing these results with those in table 2 shows, that even if the categorization performance on the category Persons is good, the framework might use mainly contextual information for classification (e.g. it uses parts of streets or buildings). Focusing on the other two categories one can recognize that SIFTs and Similarity-Measure (SM) also tend to use contextual information, whereas the moment invariants (MI) use more object relevant data. The right half of table 1 shows the results for λ_d. The following clear coherence can be seen. If a high percentage of the weighted weak hypotheses contain object data instead of contextual information (which means λ_h is high), then also the value of λ_d (meaning a new training image was classified mainly by object related information) is high.

Table 1. The measures λ_h and λ_d using various description techniques

-	λ_h				λ_d			
Data	MI (t_1)	MI (t_2)	SIFTs	SM	MI (t_1)	MI (t_2)	SIFTs	SM
Bikes	3.0	1.17	0.45	0.85	2.19	2.0	0.5	0.17
Persons	0.28	0.39	0.25	0.39	0.42	0.56	0.12	0.16
Cars	1.13	1.18	0.1	0.25	0.52	0.59	0.06	0.08

To perform useful localization with this weakly supervised system we may require $\lambda_h > 1.0$, which just means that a significant number of local descriptors

[4] The images are chosen sequentially from the database. This means we took the first 300 images of an object class and took out every second image for the test set.

Table 2. The ROC-equal-error rates of various specific combinations of region extractions and description methods on the three categories of our new dataset (MI ... moment invariants, SM ... Similarity Measure)

Data	MI (t_1)	MI (t_2)	SIFTs	SM
Bikes	72.5	76.5	76.4	74.0
Persons	81.0	77.2	70.0	74.1
Cars	67.0	70.2	68.9	56.5

Table 3. A comparison of the localization criterion by Agarwal et al. [1] with our ground truth in the first two rows. And additional for Motorbikes (100 images) of the Caltech database in the last row

Data	L(T)	L(F)	L.P.	L+Cat
Bikes	115	35	76.7	56.0
Persons	83	67	55.3	48.2
Cars	72	78	48.0	35.8
Motorbikes	96	4	96.0	88.5

is relevant for object localization. This is also supported by the observation that high values of λ_d correspond with high values of λ_h.

Table 3 shows the results (with Moment Invariants and affine invariant interest points (t_1)) achieved by comparing the localization measure of Agarwal et al. [1] with our ground truth. The first row (L(T)) shows the number of all positive test images where just the correct localization was measured, not the categorization performance. The second column (L(F)) shows the same rate for the false localizations. The third column (L.P.) shows the localization performance on the test images in percent. Note that values around 50 percent are not close to guessing, regarding that the objects cover just a small region in the images. The last column shows the result in ROC-equal error rate for categorization combined with correct localization. It can be seen that the localization performance on the category Bikes is highest, but even on Persons the performance is surprisingly high. The last row shows that localization is much easier for the simpler Caltech (motorbikes) dataset. To compare with an existing approach we mention the classification performance of 94% achieved by Leibe [9] on this dataset. Their model based approach also localizes the object, but uses high supervision in the training procedure (whereas we use almost no supervision). This is not in contradiction with the results presented in table 1. It just shows that even if a significant number of local descriptors is located in the background (low values for λ_h and λ_d), the selection of the relevant $R_{C_{max}}$ is still quite good.

Figure 4 shows examples of the localization of Bikes in test images. The bottom row shows the direct localization with the black squares representing regions with a high probability of the object location (each black square may contain several best matches for firing hypotheses). In the top row we show the effective localization where the light gray squares mark the clusters and the dark gray cross marks the final output $R_{C_{max}}$ of our Selection Algorithm. Note that we did not use ground truth for this localization. The performance of the Selection Algorithm can be shown as it finds the correct location in images with a high percentage of hypotheses firing on the object (the first two columns) as well as finding the correct location when more hypotheses fire in the background (the third column of figure 4 shows an example). In general the localization often fails when the object appears at a very low scale.

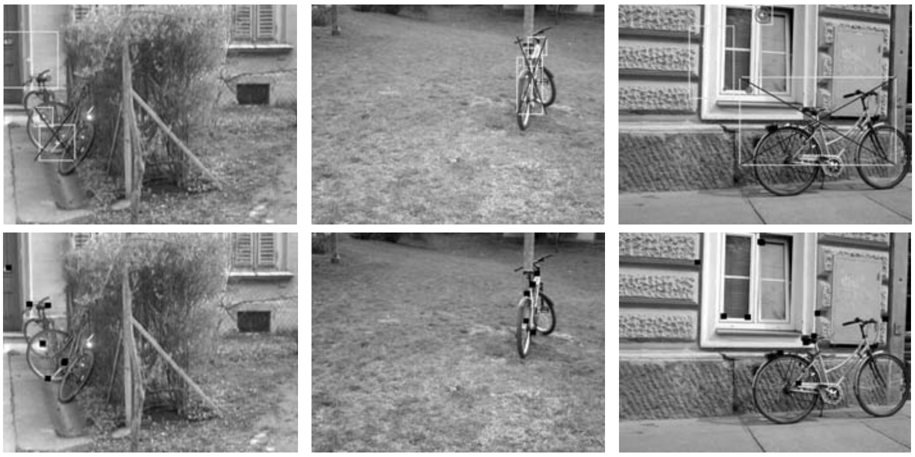

Fig. 4. Examples of the localization performance for Bikes

5 Summary and Conclusions

In summary, this work shows the first systematic evaluation of *object localization* for a weakly supervised categorization system. Supervision is regarded weak, when labeled training images are used which contain the objects of interest at arbitrary scales, poses, and positions in the images. A further important requirement is a balance of background with respect to different object categories, so that learning of context is inhibited. We have set up a very complex new image database which meets all the above requirements. We also acquired localization ground truth for all relevant objects in all images.

We have extended our categorization system [14] that calculates a large number of weak hypotheses which are based on a variety of interest operators, segmentations, and local descriptors. Learning in this system is based on Boosting. Localization measures have been defined and evaluated which are in correspondence with such a learning approach. Our 'direct' localization measures λ_h and λ_d show that even if a balanced database is used, many descriptors are still located in background regions of the images. However, the more general localization measure of Agarwal et al. [1] still yields rather good results (regarding the image complexity). Furthermore, there is a significant intra-class variability. Localization performance is class-dependent. For our database the best localization can be achieved for Bikes, and is much better than the localization for Persons and Cars. On easier datasets like e.g. motorbikes (Caltech) the localization is rather straightforward. This is because the prominency of the object reduces the complexity of a weakly supervised approach to distinguish between object and background.

An important general question might be raised: Have we already reached the frontier of categorization and localization based on local features without using any further model or supervision? We believe, that a general cognitive approach should avoid more supervision but will require more geometry. Thus, our future research will focus on the learning of sparse geometric models.

Acknlowledgements

This work was supported by the Austrian Science Foundation FWF, project S9103-N04.

References

1. S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE PAMI*, 26(11), Nov. 2004.
2. P. Carbonetto, G. Dorko, and C. Schmid. Bayesian learning for weakly supervised object classification. Technical report, INRIA Rhone-Alpes, Grenoble, France, August 2004.
3. D. Comaniciu and P. Meer. Mean shift: A robust approach towards feature space analysis. In *IEEE PAMI*, volume 24(5), pages 603–619, 2002.
4. R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proc. European Conference of Computer Vision*, pages 242–256, 2004.
5. Y. Freund and R. Schapire. A decision theoretic generalisation of online learning. *Computer and System Sciences*, 55(1):119–139, 1997.
6. M. Fussenegger, A. Opelt, A. Pinz, and P. Auer. Object recognition using segmentation for feature detection. In *Proc. ICPR*, 2004.
7. L. Van Gool, T. Moons, and D. Ungureanu. Affine / photometric invariants for planar intensity patterns. In *Proc. ECCV*, pages 642 – 651, 1996.
8. S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *In Proc. of British Machine Vision Conference*, 2004.
9. 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, Prague*, May 2004.
10. B. Leibe and B. Schiele. Scale-invariant object categorization using a scale-adaptive means-shift search. In *DAGM'04 Pattern Recognition Symposium, Tuebingen, Germany*, Aug. 2004.
11. D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999.
12. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proc. ICCV*, pages 525–531, 2001.
13. K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*, pages 128–142, 2002.
14. A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. ECCV*, pages 71–84, 2004.
15. U. Ruthishauser, D. Walther, C. Koch, and P. Perona. Is bottom.up attention useful for object recognition? In *Proc. CVPR*, 2004.
16. A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *Proc. CVPR*, 2004.