An Evaluation of an Object Recognition Schema Using Multiple Region Detectors

Meritxell Vinyals [a] Arnau Ramisa [a,1], and Ricardo Toledo [b]
[a] *IIIA, Artificial Intelligence Research Institute*
CSIC, Spanish National Research Council
[b] *CVC (Computer Vision Center).*

Abstract. Robust object recognition is one of the most challenging topics in computer vision. In the last years promising results have been obtained using local regions and descriptors to characterize and learn objects. One of these approaches is the one proposed by Lowe in [1]. In this work we compare different region detectors in the context of object recognition under different image transformations such as illumination, scale and rotation. Additionally, we propose two extensions to the original object recognition scheme: a Bayesian model that uses knowledge about region detector robustness to reject more unlikely hypotheses and a final verification process to check that all final hypotheses are coherent to each other.

Keywords. Object Recognition, Local Features, Affine Region Detectors, SIFT

1. Introduction

Since its beginnings, object recognition has been amongst the most important objectives of computer vision. One of the main issues to solve this challenge is finding new ways to represent objects that allow a reliable recognition under a wide range of variations in lightning, pose or noise. Lately, one of the most successful approaches to this problem has been the use of local feature regions to characterize and learn objects.

Local feature regions correspond to interesting elements of an image, which can be detected under larger changes in viewpoint, scale and illumination. Many different types of feature region detectors have been developed recently [1,2,3,4]. Mikolajczyk in [5] reviewed the state of the art of these affine covariant region detectors individually.

Lowe developed in [6,1] a object recognition scheme that uses SIFT points (Scale Invariant Feature Transform) to learn and recognize objects. Matches between the learned object models and the new image are computed and refined through various stages. This approach achieved good results detecting previously learned objects in cluttered environments with changes in pose and with partial occlusion.

In this work we use this scheme with the region detectors that give better results in the comparison done by Mikolajczyk et al. and the SIFT descriptor to test its performance in a object recognition task under changes in lightning, pose and scale. For

[1] Correspondence to: Arnau Ramisa, UAB Campus, 08193 Bellaterra, Spain. Tel.: +34 93 5809570; Fax: +34 93 5809661; E-mail: aramisa@iiia.csic.es

our experiments we use the well known object databases ALOI [7], COIL-100 [8] and GroundTruth100-for-COIL [9].

Additionally we propose two improvements to the original Lowe recognition scheme: a Bayesian model that calculates hypothesis probability using knowledge about the robustness of regions detectors to different transformations and a final verification process to asses that all final hypotheses are coherent to each other.

The rest of the paper is structured as follows: In Section 2 we explain the object recognition method developed by Lowe and our proposed enhancement. Then, in Section 3 we detail the experiments and provide an analysis of the results. Finally, in section 4 we discuss the conclusions and some lines of future research.

2. Object detection scheme

In this section we briefly describe the object detection scheme proposed by Lowe in [6,1] and our proposed modifications. An overview of this method can be seen in Fig. 1

2.1. Local region extraction

The first step in the object recognition scheme is to detect and describe the local interest regions in model and test images. Lowe proposed in [6,1] its detector and descriptor: the Difference-of-Gaussian detector (DoG's) and the Scale Invariant Feature Transform (SIFT) respectively. In our approach, in addition to the DoG's regions, we wanted to test the performance of the object recognition scheme with other local regions. Mikolajczyk et al. compared in [5] some of the latest affine-covariant region detectors. Based on this comparison we have chosen the three region detectors that obtained better results: the Harris-Affine [2], the Hessian-Affine [2] and the MSER (Maximally Stable Extremal Regions) [3]. We use all these detectors combined or separated to extract the different interest image regions.

In order to match different occurrences of an interest region it is necessary to use a local descriptor to characterize it. In this work , as in the original scheme, we have used the SIFT descriptor. This descriptor divides the local region into several sub-regions and computes histograms of the orientations of the gradient for every sub-region. The values of all bins of the histograms are then concatenated, forming a descriptor vector of 128 dimensions.

2.2. Descriptor matching

Here we explain the descriptor matching process used to identify different object instances in test images by matching image descriptors to an object descriptor database that stores the object models.

Descriptors from a test image are matched to descriptors stored in the database using Euclidean distance. Each new local descriptor is matched against its nearest neighbour in the model database. Then, the second nearest neighbour is used to decide if the match is valid or if it is a false correspondence: if the first and the second nearest neighbours are very close, the match is considered incorrect. Namely:

$$\frac{NN_2}{NN_1} > 0.8, \tag{1}$$

Where NN_1 is the distance to the first nearest neighbour, NN_2 is the distance to the second nearest neighbour, and 0.8 is a threshold value determined experimentally by Lowe [1]. Descriptors are efficiently matched using a k-d tree structure and the Best-Bin-First algorithm. As a result of this process, a set of matches between models and the test image is found. These matches are the first hypotheses about which objects appear in the new image. An example of these preliminary matches can be seen in Fig.2.

2.3. Clustering and pose estimation algorithms

In this stage the scheme combines different clustering and pose estimation algorithms to find consistent sets of descriptor matches within the initial matches set and give an estimation of the transformation occurred. We propose two modifications to the original scheme: a bayesian model combined with the RANSAC algorithm to consider hypothesis probabilities given a transformation and a process of verification of the final hypotheses.

Typically, the initial set of matches coming from the descriptor matching process is still contaminated with correspondences that come from other objects or background texture. To discard most of the false matches and distinguish between different object instances, the next step is a clustering with a generalised Hough transform. In this step, the matches belonging to the same model are clustered according to its scale, position and orientation. Each cluster with three or more matches is an hypothesis and is subject to further verification. Since each match votes for more than one bin (the selected bin and its adjacent) matches may appear in more than one hypothesis. These repeated matches

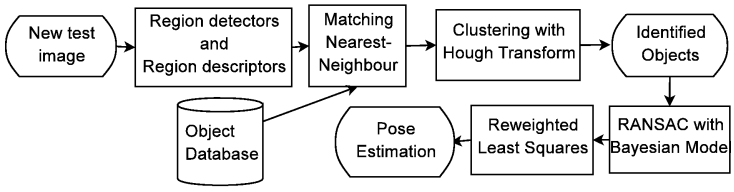


Figure 1. Diagram of the detection scheme.

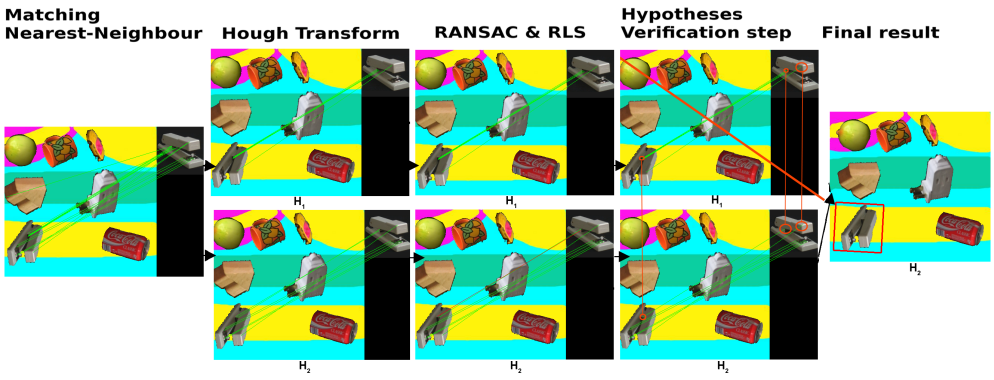


Figure 2. Example of execution where the final hypotheses verification step detects two hypotheses that refers to the same object instance

will be used in the last verification hypothesis step to detect hypothesis that refer to the same object instance and keep only the one which has higher robustness and support.

Next, for all clusters with three or more correspondences, the RANSAC algorithm is used to obtain an estimation of the object pose and identify the remaining outlier if any (see Fig.2 for an example of outliers, displayed as red lines, detected by RANSAC). Although typically the model parameters estimated by RANSAC are not very precise, it is used due to its high tolerance to outliers (it has a breakdown of 50%). Finally, RANSAC best hypotheses inliers are used in the reweighted least squares method (IRLS) to recompute the estimated model parameters accurately.

2.3.1. RANSAC with Bayesian Model

In this section we propose a modification over RANSAC to improve pose estimation results. In the pose estimation process sometimes the correct hypothesis is discarded in favour of a more supported, yet improbable, hypothesis. We define an improbable hypothesis as one that proposes a transformation where detectors are know to have very low repeatability rates. However we observe that given a number of matches between an object and an instance not all pose estimation hypotheses are feasible or equiprobable. Typically RANSAC returns as best hypothesis the one that maximizes the number of inliers or the hypothesis with the least median residual. In our approach we propose to modify these functions to consider not only that the hypothesis gets support from the input data set but also the hypothesis probability given that data set (Eq. 2).

$$V_H* = V_H \cdot (1 + P(\bar{H} \mid |D|)) \tag{2}$$

where V_H is the typical cost function of RANSAC that calculates how good is one hypothesis, H is the transformation estimation hypothesis and $|D|$ is the cardinality of the set of matches that support it.

We propose a bayesian model to calculate the probability of an hypothesis given a set of matches which uses as prior probabilities the expected detector repeatability rates under different transformations,

$$P(H \mid |D|) = \frac{P(H) \cdot P(|D| \mid H)}{P(|D|)} \tag{3}$$

where H is the pose estimation hypothesis and $|D|$ is the cardinality of the set matches that support H. We consider $P(|D|)$ equiprobable given a model and we set its value to $\frac{1}{|D^m|}$ where $|D^m|$ is the number of descriptors contained in the model. We also define all the hypotheses space as equiprobable ($P(H)$).
As the object recognition system can use more than one region detector with different robustness and capabilities, we define $P(|D| \mid H)$ in a more general form where the probabilities are calculated for each detector and weighted by its presence in the descriptor matches set:

$$P(|D| \mid H) = \sum_{i \in detectors} p_i \cdot P(|D_i| \mid H), \tag{4}$$

where p_i is the percentage of matches with regions type i (notice that $\sum_{i \in detectors} p_i = 1$), $|D_i|$ is the cardinality of the set of matches using the region detector i and $detectors$ is the set of all detectors used in the extraction of image characteristic regions.

Given an hypothesis of the transformation occurred, the probability that the system retrieves a number of matches depends on the robustness of the detectors to that transformation. Hence we propose to define $P(|D| \mid H)$ in function of the results of the experiments described in detail in Section 3 (see Fig.4 and Fig. 5 for experiment results). The repeatability rate expected for each detector is obtained interpolating the result value of the two closer sampled points of the transformation space in our experiments. The hypothesis probability distribution is modeled using a normal distribution with a mean equal to the number of matches we expect to have (the product between the percentage expected and the total number of regions from the model image) and a variance equal to a half of the mean. The final hypothesis probability is obtained by adding the hypothesis probability calculated for each detector separately weighted by the percentage of the regions extracted with that detector in the input data.

2.4. Final hypotheses verification step

In our experiments we observe that because of in the Hough transform matches can be duplicated in different clusters, final hypotheses can present a non-disjoint set of support data (some repeated matches). To solve this, it is not acceptable to keep just one hypothesis and discard the rest, because we would not consider that an object can have more than one instance in the image. However, since a descriptor match can belong to only one object instance, we can assume that if the inlier data sets of different hypotheses are not disjoint then they must refer to the same object or only one of the hypotheses can be the correct. In that case we propose to keep the hypothesis with the highest number of inliers or, if several hypotheses have an equal number of inliers, the one with less transformation error. To illustrate this process, see Fig. 2 where the Hough Transform ends with two clusters of valid matches from the initial set. These both hypotheses reach the final verification step with different pose estimations but since they have matches in common (displayed as lines between the two images) the scheme detects that they refer to the same object instance and discard the hypothesis with less number of matches.

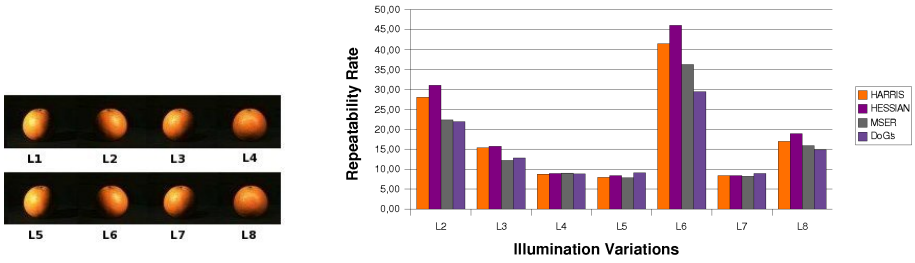
3. Experimental Results

First we explain the experiments designed to assess the robustness and capabilities of the different region detectors and the descriptor used. The empirical results obtained in this test are used as prior probabilities in the Bayesian Model. Secondly we present the experimental results for our object recognition and pose estimation scheme by evaluating its performance with public image databases.

3.1. Region Repeatability Results

As explained in [10] one of the characteristics to measure the performance of a good region detector is repeatability. We evaluate the repeatability rates of each detector under three different transformations: illumination variation, scale change and image rotation. The measure of repeatability takes into account the uncertainty of detection. A point x_i detected in image I_i is repeated in image I_j if the corresponding point x_j is detected in image I_j where x_j is defined as:

$$\{x_j\} = \{x_i \mid T_i \cdot x_i \in I_j\} \tag{5}$$



(a) Object data set

(b) Illumination repeatability results

Figure 3. Illumination test

A repeated point is in general not detected exactly at position x_j, but rather in some neighbourhood of x_j. The size of this neighbourhood is denoted by e and repeatability within this neighbourhood is called e-repeatability. The set of points pairs (x_i, x_j) which correspond within an e-neighbourhood is defined by :

$$Rj(e) = \{(x_i, x_j) \mid dist(T_i \cdot x_i, x_j) < e\} \tag{6}$$

We set parameter $e = 1.5$, as is proposed in [10] for all experiments.

3.1.1. Illumination Variation

In this experiment we evaluate the repeatability rate of each region detector for a set of scenes where an object is presented under different illumination conditions. We use the ALOI (Amsterdam Library of Object Images) image dataset which provide one-thousand small objects recorded under 24 different illumination conditions (Figure 3(a)).

Fig. 3(b) depicts the means over the tests done with 100 different objects and their respective 24 images grouped by illumination variation. The repeatability rate varies considerably among different illumination changes. Observe that although different detectors generally present similar results, significant differences appear in tests with soft illumination changes (L2,L6). In these cases Hessian and Harris detectors produce considerably better results than MSER and DOG's.

3.1.2. Scale Changes

In this experiment we assess the repeatability rate of the detectors under 10 different scales (1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.25 and 3.5).
Fig. 4 depicts means over the 50 tests run using objects from the ALOI image database. Observe that significant differences appear among different detectors. DoG's detector produces better results than all other detectors keeping their repeatability rate less affected among the scale variations applied. MSER detector achieves higher rates than Hessian and Harris, these last two reporting nearly identical results.

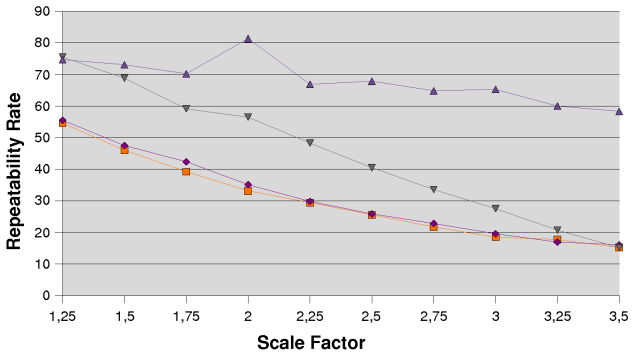


Figure 4. Scale repeatability results

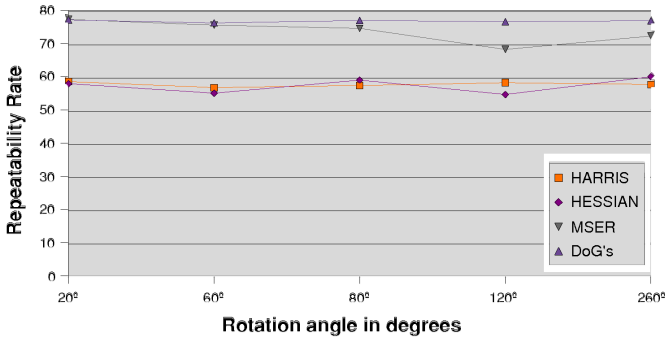


Figure 5. Rotation repetability results

3.1.3. Image Rotation

In this experiment we evaluate the repeatability rate of the detectors under different rotation angles. We report means over 50 tests run using images of objects from the ALOI image database. Images are rotated at 6 different angles: $0\,^{\circ}$, $20\,^{\circ}$, $60\,^{\circ}$, $80\,^{\circ}$, $120\,^{\circ}$ and $260\,^{\circ}$. As you can see in Fig. 5 the detector repeatability rates are independent of the rotation angle applied producing similar results among all rotation angles. However rates obtained when some rotation is applied to the image varies among detectors. While DoG's and MSER produce results close to 80%, Harris and Hessian present a repeatability rate of only 60%.

3.2. Number of regions extracted by detector

In this experiment we aim to compare the number of regions extracted by each detector in 10 different image resolutions. Usually objects represent small regions in images, therefore detectors that retrieve few regions can have problems in order to describe an

object present in a frame. The results reported in Fig. 6 are the means over 50 tests run over different images. Observe that we have a linear relationship between the number of regions extracted by each detector and the image resolution. Furthermore significant differences appear among the number of regions extracted by each detector: DoG's always find the highest number of regions, followed by Hessian, Harris and MSER.

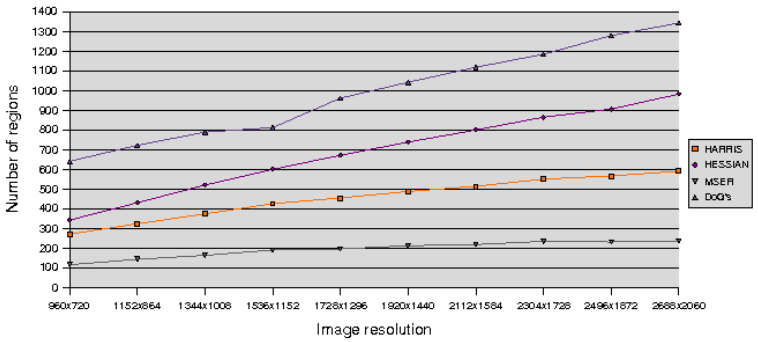


Figure 6. Number of regions extracted per image resolution

3.3. Object recognition and pose estimation experiments

In this experiment we aim to assess the performance of the scheme described in section 2 in the identification of which objects are present in an image scene (object recognition) and which is its transformation with respect to the model (pose estimation). We use the GroundTruth100-for-COIL image database which is composed of 100 images each one with different objects instances from the COIL-100 object image database. Additionally, this database includes information about which objects appear in each image and the scale change and rotation angle applied in each case.

Fig. 7 (left) depicts the percentage of objects correctly detected using each single detector and all detectors combined. Furthermore the figure reports the percentage of correct identified objects for which the object recognition scheme has been capable of generating an estimation of the geometric transformation occurred (a minimum of matches are required in each step of the scheme to calculate the pose estimation).

As you can see in Fig. 7 (right) DoG's produces the highest rate of correct identified objects (93%) , followed by Harris and Hessian detectors (38% and 43% respectively). MSER detector reports the worst results with a very poor percentage of correct matches (13%). It also shows that when all detectors are used to detect object regions the final number of matches increases (99%). These results are correlated with the number of regions found by each detector rather than the repeatability rates reported in our experiments (see section 3.1). Since the image resolution used is quite low, and consequently regions corresponding to objects have small sizes, detectors that extract fewer numbers of regions are likely to have low performances since they detect very few regions or none for each object. Although we could have used images with higher resolutions usually object recognition applications (robot navigation, video surveillance ...) require to work

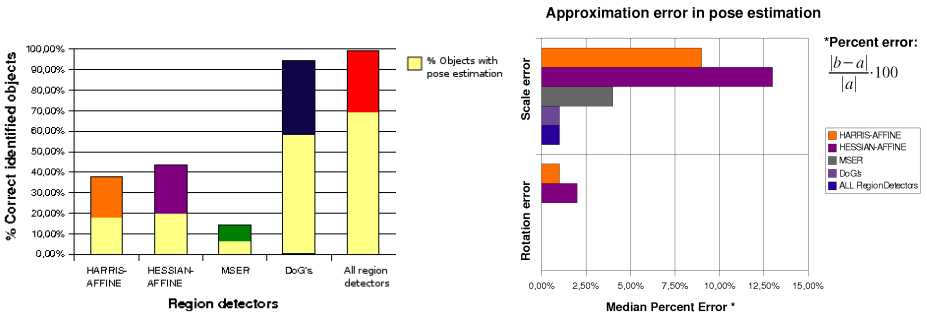


Figure 7. Object recognition results grouped by region detectors.

with limited image sizes. Hence, a suitable detector for object recognition is usually required to achieve good results in poor quality images. Consequently DoG detector is more suitable to be used in that conditions than other detectors like MSER which also present good repeatibily rates. We can conclude that when using a recognition object scheme with high tolerance to outliers best performances are achieved by detectors that find higher number of regions (although they have more false matches) than very reliable detectors that give fewer matches.
In that experiment we also measure the accuracy of the pose estimation provided by the object recognition scheme. The results (see Fig. 7) show the median percent error produced in the estimation of the scale change and rotation occurred between the model and the instance. These results match with the scale and rotation repeatability rates reported in our experiments (see Section 3.2) since matches over DoG's and MSER regions allow more accurate scale and rotation approximation than the ones obtained with Harris and Hessian regions. Finally we also observe that when we use all detectors the results are quite similar to the ones produced using only DoG's regions due to the higher number of DoG's regions compared to other detectors.

4. Conclusions

In this work we have evaluated the performance of various state-of-the-art region detectors in the Lowe objection recognition scheme. According to our experiments, region detectors that find a higher number of regions obtain better results in object recognition tasks. From these results we also observe that detectors achieve different performances under different kind of transformations. We also conclude that in order to choose a descriptor is not only important its repeatability also the number of regions that it extracts from the image. Very reliable detectors with high levels of repeatability are not suitable for object description because they extract very few regions per image. Finally we observe that the combination of different region detectors improves object recognition results. Furthermore, we propose two modifications to the original scheme: a bayesian model that uses region detector robustness as prior knowledge to reject improbable transformations and a process to verify final hypothesis. Our work argues in favour of researching how to combine region detectors taking into account the information about its

robustness under different transformations. Finally, as future work, a formal comparison of this new approach with respect to the original object recognition scheme should be provided.

Acknowledgements

This work has been partially funded by the European Social Fund, the MID-CBR project grant TIN2006-15140- C03-01, TIN 2006-15308-C02-02 and FEDER funds. The work of Meritxell Vinyals is supported by the Ministry of Education of Spain (FPU grant AP2006-04636) whereas the work of Arnau Ramisa is supported by the FI grant from the Generalitat de Catalunya.

References

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *Int. J. Comput. Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[3] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions." in *Proceedings of the British Machine Vision Conference 2002, BMVC 2002, Cardiff, UK, 2-5 September 2002*. British Machine Vision Association, 2002.

[4] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector." in *Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I*, ser. Lecture Notes in Computer Science, vol. 3021. Springer, 2004, pp. 228–241.

[5] K. Mikolajczyk, *et al.*, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 2, pp. 43–72, 2005.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV '99: Proceedings of the International Conference on Computer Vision-Volume 2*. Washington, DC, USA: IEEE Computer Society, 1999, p. 1150.

[7] I. S. I. Systems, "Aloi amsterdam library of object images."

[8] S. K. N. Sameer A. Nene and H. Murase, "Coil-100 columbia object image library." [Online]. Available: http://www1.cs.columbia.edu/CAVE/publications/pdfs/Nene_TR96_2.pdf

[9] T. of Vision (ITC-irst), "Groundtruth100-for-coil object image database." [Online]. Available: http://tev.itc.it/DATABASES/objects.html

[10] C. Schmid, R. Mohr, and C. Bauckhage, "Evaluation of Interest Point Detectors," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 151–172, 2000.