Saliency of Interest Points under Scale Changes

Daniela Hall, Bastian Leibe and Bernt Schiele
Perceptual Computing and Computer Vision Group, ETH Zurich
Haldeneggsteig 4, CH-8092 Zurich, Switzerland
{halld,leibe,schiele}@inf.ethz.ch

Abstract

Interest point detectors are commonly employed to reduce the amount of data to be processed. The ideal interest point detector would robustly select those features which are most appropriate or salient for the application and data at hand. There is however a tradeoff between the robustness and the discriminance of the selected features. Whereas robustness in terms of repeatability is relatively well explored, the discriminance of interest points is rarely discussed. This paper formalizes the notion of saliency and evaluates three state-of-the-art interest point detectors with respect to their capability of selecting salient image features in two recognition settings.

1 Introduction

Computing feature correspondence under scale changes is necessary for many computer vision applications. For example, this is the case for image retrieval, recognition from local features, stereo vision, image mosaicking, and 3D modeling from wide baseline images. Interest point detectors are often used in order to reduce the amount of data and to compute feature correspondences for interest points only [15]. The repeatability of interest points detectors is therefore important and has been studied quite intensively [9]. Generally however, repeatability does not say anything about the probability that an image feature can be matched correctly, that is that the image feature is salient for the current application. This article evaluates interest point detectors with respect to the saliency of the selected features. Recently, scale invariant interest point detectors have been described. Therefore it is interesting to explore the behavior of such interest point detectors under scale changes.

Saliency of an image feature can be defined to be inversely proportional to the probability of occurrence of that image feature. That is the lower the probability of occurrence the higher the probability of a correct match and therefore the higher the saliency or discriminance of that feature. Selecting only salient features by means of an appropriate interest point detector has the potential to improve the overall result of the matching as well as to reduce computation time.

Section 2 defines the saliency of an image feature and proposes a saliency measure. Section 3 introduces three popular and state-of-the-art interest point detectors for preprocessing. Obviously, the saliency of image features depends on the feature vectors used. It does, however, also depend on the application and the data. Section 4 therefore reviews the feature description and matching technique used in the subsequent experiments. Section 5 evaluates the interest point detectors with respect to the selection of salient features. The validity of the approach is demonstrated on the example of two image databases.

2 Salient Features

The most salient or discriminant image features are those that allow to distinguish one feature from others. An image feature that is present in only a single image or on a single object would allow to distinguish this image or object from all others. According to the definition, such a feature has maximum saliency. The term "salient feature" has previously been used by many other researchers, for example [2, 13, 16, 17], although definitions vary.

Intuitively, saliency corresponds to the rarity of a feature. Following Walker et al [17], we want to formalize this by defining the saliency over the probability density in feature space. In their work, they approximate the probability density by a mixture of Gaussians. This permits efficient use of the resulting saliency estimate, but the accuracy of the approximation is naturally limited by the Gaussians' capability to model the true probability distribution. Since our focus is not on efficiency, but on accurately measuring the saliency of points obtained by existing state-of-the-art methods, we prefer a more exact approximation by directly using all available training features. Given a feature space populated by N learned points, the probability density in a certain volume V containing K points can be estimated by $p(x) = \frac{K}{NV}$. The saliency is inversely proportional to this density. For the estimation of the saliency s of an image feature f, we consider a Parzen window in form of a hypersphere V_ε around f. This leads to the following measure:

$$s(f, \varepsilon) = \frac{1}{p} = \frac{NV_\varepsilon}{K} \tag{1}$$

In the experiments reported below, N and ε are constant. The saliency therefore becomes

$$s_\varepsilon(f) \approx \frac{1}{K} \tag{2}$$

A small number of samples in the considered ε-sphere indicates a high saliency, since the probability of confusion is small. The equations above allow to determine the saliency of features: $s_\varepsilon(f)$ is computed by counting the elements K within a sphere of radius ε centered on the feature vector of f. The robustness is increased when several image features are considered. Using the measure to calculate the saliency for all available image points, however, is computationally prohibitive. The determination of saliency requires a search in descriptor space, which has the same computational complexity as calculating feature correspondence itself. We therefore need a means to select image points.

3 Interest Point Detectors

An idea to reduce the computational cost for image analysis is to take the image content into account by pre-screening with an interest point detector. In the following we introduce three state-of-the-art interest point detectors, which have been shown to yield high repeatability [9, 15]. The next section evaluates those interest point detectors with respect to the saliency of the selected points.

Harris Corner and Edge Detector. The Harris corner and edge detector [5] uses the auto-correlation function to determine locations where the signal changes in two dimensions. The following matrix A, related to the auto-correlation function, is computed which contains first derivatives, L_x, L_y, of the image. The eigenvectors of A give the principal curvature of the auto-correlation function. Two high eigenvalues indicate a detection point. The corner response function R is used for point detection without explicitly computing the eigenvalues. R is based on the determinant and the trace of A. The factor α determines the maximum ratio of eigenvalues for which R is positive. We use $\alpha = 0.04$ as suggested by Harris.

$$R = \det(A) - \alpha \operatorname{trace}^2(A), \text{ with } A = G(\sigma) * \begin{pmatrix} L_x^2 & L_x L_y \\ L_x L_y & L_y^2 \end{pmatrix} \tag{3}$$

Lindeberg Interest Point Detector. An alternative method is based on interest points proposed by Lindeberg [7]. Lindeberg generalizes interest points to scale space. He defines a normalized scale-space extremum of a differential entity $\mathcal{D}_{\mathrm{norm}}L$ as a point in scale space that is simultaneously a local extremum with respect to the spatial domain and the scale parameter. In terms of derivatives, such points satisfy

$$(\partial_\sigma (\mathcal{D}_{\mathrm{norm}}L))(\vec{x}; \sigma) = 0, \qquad (\nabla(\mathcal{D}_{\mathrm{norm}}L))(\vec{x}; \sigma) = 0. \tag{4}$$

Since scale space and spatial domain are discrete, the interest point detection is implemented by searching an extremum in three dimensions (x, y, and scale). In our system, the intrinsic scale of a feature is computed from the normalized Laplacian. The Laplacian has been chosen due to its blob detection property. Choosing the Laplacian also as normalized differential entity for interest point detection has several advantages. Firstly, the blob detection property is conserved, and, secondly, the data computed during the feature description can be reused (see Section 4). This means, that the interest points can be computed with little additional computing cost. We point out that other entities can be used, which will result in different types of interest points.

Harris-Laplacian Interest Point Detector. Recently, Mikolajczyk and Schmid [9] have proposed a new interest point detector that combines the repeatability of the Harris detector with the scale invariance of the Lindeberg detector. This Harris-Laplacian interest point detector is computed by a similar principle as the Lindeberg detector, that is interest points are searched in the volume spanned by the spatial coordinates and the scale parameter (x,y, and scale).

In a first step, a scale space representation of the Harris function F_H is built. At each level of the scale space, we search local maxima of the Harris function

$$F_H(\vec{x}, \sigma_n) > F_H(\vec{x_w}, \sigma_n), \forall \vec{x_w} \in W \tag{5}$$

where W is the 8 point neighborhood of $\vec{x}$. For each of these candidate points on the different levels, we then test if the candidate point simultaneously presents a maximum over scale by evaluating the normalized Laplacian in scale direction. Only points that fulfill both conditions are returned.
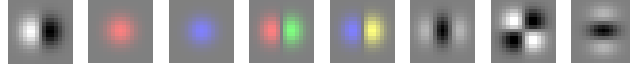
Figure 1: Feature description vector based on Gaussian derivatives. G_x^Y, G^{C_1}, G^{C_2}, $G_x^{C_1}$, $G_x^{C_2}$, G_{xx}^Y, G_{xy}^Y, G_{yy}^Y of size $\sigma = 2.3$ oriented to $\theta = 0$.

4 Application to Feature Matching

Our goal is the evaluation of interest point detectors with respect to their capability of choosing salient points. Such an evaluation is only possible in the context of an application. Our test case is image indexing under scale changes. The following section describes the type of features we use and our implementation of the saliency computation.

4.1 Description of Local Image Features

Feature descriptions by Gaussian derivative operators have become popular in the recognition community [8, 9, 11, 14, 15]. A local feature (represented by a small image window) is described by its projection on a particular set of Gaussian derivatives. Gaussian derivatives of low orders measure the basic geometries of local image features [6]. In addition, they can be normalized for local scale and orientation, such that the description is invariant to those influences. The discriminance of feature vectors can be increased by extending them to the color domain. Among the existing color spaces, the luminance chrominance space has been found to be the most adapted for feature description [3, 4]. In the experiments, we use the Gaussian derivative vector displayed in figure 1 which has produced good results in previous experiments [4].

In order to consider an image feature at its appropriate scale, we apply Lindeberg's automatic scale selection algorithm [7]. For all image features, it scans the scale space and returns an intrinsic scale. This intrinsic scale can be reliably retrieved even when the input image undergoes scale changes [7, 9]. By scaling the Gaussian derivative vector to its intrinsic scale, we thus obtain a scale invariant feature description. The resulting feature vectors populate a multi-dimensional feature space. In this particular space, the distance between feature vectors can be used as a measure for similarity of features in image space, as has been observed by Ohba and Ikeuchi [10]. Retrieval of image features can therefore be described as a nearest neighbor problem.

4.2 Saliency Computation and Matching

All appearance based recognition systems require a learning phase prior to recognition. In our system, learning is performed by projecting a dense grid of overlapping neighborhoods onto the (scale and orientation normalized) Gaussian derivative vector. The feature vectors need then to be stored appropriately. For efficient retrieval, the type of indexing structure is crucial. In our system we use a hash tree structure similar to the one used in [1, 15]. The tree is based on the principle of KDB-trees [12] that all major axes of the descriptor space are divided subsequently into n parts of equal length. This tree has the property that the feature vector is used as direct access key to the leaf containing the most similar model feature vectors.

For feature matching, a local feature is selected from an unknown image and projected onto the Gaussian derivative vector normalized to scale and orientation. Matching candidates are determined by searching the feature description space that is populated by feature vectors acquired during learning. All feature vectors situated within a certain distance in descriptor space from the observed feature vector are considered as matching candidates. The most likely object is obtained by processing the candidates' image labels.

5 Performance of Selection Strategies

In this section, we evaluate the interest point detectors of section 3 with respect to the saliency of their results in the context of image indexing under scale changes.

5.1 Influence of the Detector Parameters

The interest point detectors presented in section 3 each depend on a set of parameters, which can be used to tune their behavior. Most notable is the final threshold t, which controls the quality of the returned points. For a fair comparison, we therefore evaluate the detectors over their respective parameter range in terms of precision and recall.

In a first experiment, we compare the influence of different parameter choices for each of the three detectors. For every detector, we vary the final threshold. The remaining parameters are kept at a fixed setting which has proven best in previous experiments.

As a testbed, we choose two image databases. The first database consists of 49 images of manufactured objects, such as toy cars, toy animals, different containers, cups and cutlery. The second database contains 60 images from the Corel database depicting nature scenes such as mountains, lakes, forests and wildlife. For every image pixel from the training set, we calculate a feature vector at the estimated intrinsic scale. All of these vectors are collected and stored in a tree structure. For testing, we apply the different interest point detectors to the test images. For every interest point, we calculate the corresponding feature vector and look up its saliency according to definition (2), that is the number of trained feature vectors in its ε-neighborhood.

Figure 2 shows the results of this experiment on the manufactured objects for $s \geq \frac{1}{K} = \frac{1}{100}$. The diagrams depict the precision/recall tradeoff for different threshold settings for all three detectors, averaged over the 49 test images. It can be seen that the Harris-Laplacian detector (Ha-Lap) returns only very few, but highly salient points. The Harris detector yields significantly more points than Ha-Lap even with the lowest threshold setting, but a lower percentage of them is salient. Still, the total number of salient points is higher for Harris (see Figure 2(right)). The Lindeberg detector, finally, can be tuned to yield either high precision or high recall, but not both at the same time. For other values of K, the curves shift, but the same qualitative behavior can be observed.

As saliency depends on the application and thus on the data, we performed a second experiment using a different data set for comparison. Figure 3 shows the precision and recall curves for the images of the Corel database. It can be seen from the diagrams that, due to the higher variation in image content, the returned interest points have a higher overall saliency. However, the interest point detectors exhibit the same qualitative behavior: Harris-Laplace has high precision, but low recall; Harris has slightly lower precision, but significantly higher recall; Lindeberg can be tuned over the full range.
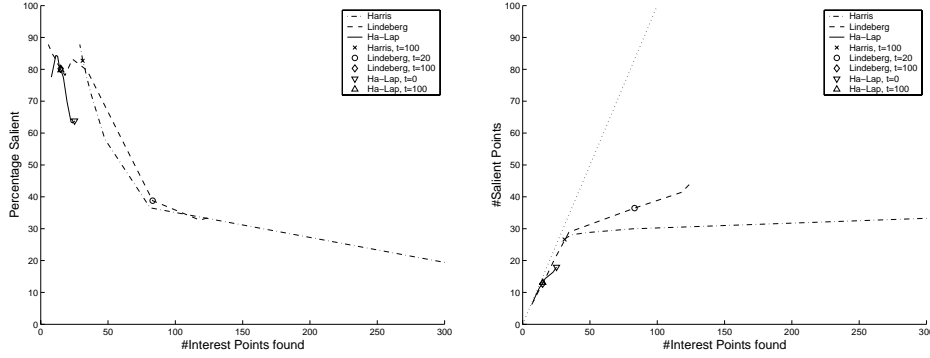
Figure 2: *Percentage (left) and absolute number (right) of salient points for three interest point detectors on 49 images of manufactured objects ($s \geq \frac{1}{K} = \frac{1}{100}$). The detectors can be tuned towards different behaviors by modifying their parameters.*
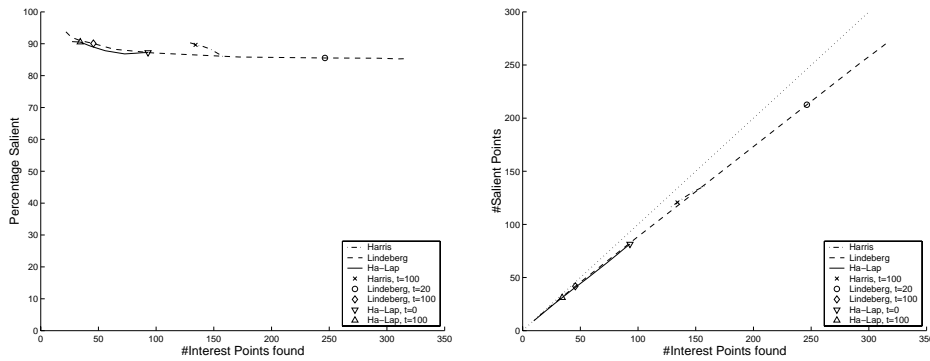

Figure 3: *Percentage (left) and absolute number (right) of salient points ($s \geq \frac{1}{K} = \frac{1}{100}$) for 60 nature images.*

5.2   Experiments on image indexing

We applied a retrieval algorithm based on probabilistic voting on both databases. To explore the behavior of the interest point detectors under scale changes, we consider images at three different scales ($1/\sqrt{2}$, 1.0, and $\sqrt{2}$) which were obtained by downsampling a large image to 47%, 33%, and 23% of its size. We use the intermediate scale for learning and all three scales for testing.

The feature description is normalized to an intrinsic feature scale which is selected within a predefined range. During learning this scale range is set to $(0.5 * 1.02^{50}, 0.5 * 1.02^{146}) \approx (1.35, 9.01)$ with a step size of 1.02^3 between levels. To allow matching invariant to scale changes, the automatic scale selection algorithm must be able to find the corresponding intrinsic scale of the unknown image. For this reason, the range of tested scales during recognition must be extended by the maximum expected scale change. In our case, this leads to a scale range from $(0.5 * 1.02^{32}, 0.5 * 1.02^{163}) \approx (0.94, 12.61)$
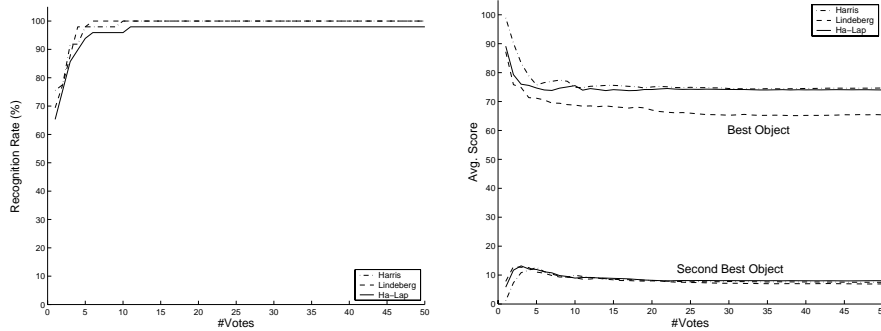
Figure 4: *Left: Percentage of objects recognized out of 49 manufactured objects with scale change* $1/\sqrt{2}$. *Right: The upper curve shows the average probabilistic voting score over number of votes. The lower curve displays the average score for the object ranked second.*

during recognition. To reduce boundary effects, we decided to restrict the range for the small images to $(0.94, 9.01)$.

We use a multi-step voting algorithm. On the lowest level, each point votes for the object that is represented most often within the considered ε-sphere. Each vote is weighted by the resulting probability $p(O_j|x) = \frac{K_j}{N_\varepsilon}$ with K_j the number of elements of the winner object O_j and N_ε the total number of elements in V_ε. On the next higher level, an object is recognized when its probabilistic voting score V_j is higher than for any other object.

$$V_j(x_1 x_2 \ldots x_n) > V_k(x_1 x_2 \ldots x_n), \forall k \neq j \tag{6}$$

with

$$V_j(x_1 x_2 \ldots x_n) = \frac{1}{n} \sum_{i=1}^{n} p(O_j|x_i) \tag{7}$$

The following recognition results are observed. All images except one are correctly identified with a small numbers of votes (between 7 to 20). This shows that the adaption of the scale range enables identification of images under scale changes. The experiment using features detected by the Harris detector requires some more votes to obtain the highest recognition (see Figure 4 (left)). This is because points selected by Harris are less salient than points selected by the other interest point detectors.

Figure 4 (right) displays the average voting score $V_j(x_1 x_2 \ldots x_n)$ for different numbers n of votes. All curves lie between 60% to 80%. Such values are sufficient for reliable recognition and the differences between the detectors are not significant. In order to get information about the reliability of the results we collected the average probabilistic voting score of the object ranked in second place (see Figure 4 (right)). This curve lies constantly below 15%.

As a consequence, all interest point detectors are equally adapted to provide a solution to the problem of image indexing under scale changes. Since the Lindeberg and the Harris-Laplacian interest point detector require significant computation effort, a general guideline would be to use the Harris detector. A more detailed interpretation of some interesting effects are given in the following section.

5.3 Interpretation of example images

In this section we interpret the behavior of the different interest point detectors in several examples. This section summarizes the properties of the different detectors when presented to difficult images.

Figure 5 (a) illustrates the point types that are preferred by the different detectors. Harris prefers corners and edges, Lindeberg selects point that are centered within a blob feature, and Harris-Laplace selects Harris points that present a maximum over scales.

The example in Figure 5 (b) displays the behavior which we identified for the largest portion of the images. The Harris-Laplacian detects significantly fewer points than Harris and Lindeberg, but in all three cases a sufficient number of salient points are detected to enable recognition.

Figure 5 (c) shows a problem case. Lindeberg detects points only on the exterior of the object. The scale of the feature is such that the border of the object is captured, but this information is not sufficient to provide identification. Identification is especially difficult since there are two other white cups in the database. Harris selects points that are situated on the only points displaying structure sufficient for identification. The Harris-Laplacian selects a small number of points which enable identification with a high probability. This example illustrates the property of the Harris-Laplacian to select only a small number of highly salient points.

Figure 5 (d) shows a counter example. Again, the Harris-Laplacian selects very few points, but in this case identification fails. As a general rule of thumb, we require a minimum number of 10 images features per object. Harris and Lindeberg both select sufficient image features and succeed in identifying the object.

The hardest cases in our database are specular objects such as the knife shown in Figure 5 (e). Harris is the only detector which selects points that lead to recognition. Lindeberg identifies the object by simple majority voting. Harris-Laplace fails due to missing image points.

From these examples we can deduct the following guideline. The Harris-Laplacian detector selects a small number of highly salient points. If this number is sufficiently high, the Harris-Laplacian produces the best results. The inconvenience is that the computation of the Harris-Laplacian is computationally expensive. The Harris and the Lindeberg detector in general select a sufficient number of salient points. Considering the lower overall saliency of these points, the Harris detector is the second best choice because the computation time is lower than for the Lindeberg detector.

6 Conclusion

We formalized a definition of saliency based on the probability density in feature space and evaluated several state-of-the-art interest point detectors with respect to this definition. Our experiments show that the detectors exhibit significantly different behavior in terms of precision and recall of salient points. We found that the Harris-Laplacian detector selects in general a small number of points which are in turn highly salient. This high saliency is produced by the requirements of the detector: a point is selected when it is a Harris point in the spatial domain and a maximum over scale, which is the reason for the selection of salient points under scale changes.

The experiments show that image indexing under scale changes is successful when considering points selected by either of the detectors. Interpreting in detail a few problem cases lets us identify the following conclusions. Probabilistic voting using image features selected by the Harris-Laplacian achieves the highest average voting score. However, in cases where there are too few selected points, identification fails. In those cases, the Harris or the Lindeberg detector can provide additional points for recognition by relaxing one of the requirements of the Harris-Laplacian.

References

[1] V. Colin de Verdière and J.L. Crowley. Visual recognition using local appearance. In *ECCV98*, pages 640–654, Freiburg, June 1998.

[2] P.J. Flynn. Saliencies and symmetries: Toward 3d object recognition from large model databases. In *CVPR'92*, pages 322–327, 1992.

[3] J.-M. Geusebroek, R. van den Boomgaard, A.W.M. Smeulders, and A. Dev. Color and scale: The spatial structure of color images. In *ECCV00*, pages I.331–341, July 2000.

[4] D. Hall, V. Colin de Verdière, and J.L. Crowley. Object recognition using coloured receptive fields. In *ECCV00*, Dublin, Ireland, June 2000.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, 1988.

[6] J.J. Koenderink and A.J. van Doorn. Representation of local geometry in the visual system. *Biological Cybernetics*, pages 367–375, 1987.

[7] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2):79–116, 1998.

[8] D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV99*, pages 1150–1157, 1999.

[9] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV01*, pages 525–531, Vancouver, Canada, July 2001.

[10] K. Ohba and K. Ikeuchi. Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects. *TPAMI*, 19(9):1043–1048, September 1997.

[11] R.P.N. Rao and D.H. Ballard. An active vision architecture based on iconic representations. *Artificial Intelligence*, 78(1–2):461–505, 1995.

[12] J.T. Robinson. The k-d-b-tree: A search structure for large multidimensional dynamic indexes. *Transactions of the Association for Computing Machinery*, 1981.

[13] B. Schiele and J. L. Crowley. Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR96*, Vienna, Austria, 1996.

[14] B. Schiele and J.L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *IJCV*, 36(1):31–50, January 2000.

[15] C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *TPAMI*, 1997.

[16] N. Sebe and M.S. Lew. Salient points for content-based retrieval. In *BMVC'01*, pages 401–410, 2001.

[17] K. N. Walker, T.F. Cootes, and Chris Taylor. Locating salient object features. In *BMVC'98*, pages 557–566, 1998.
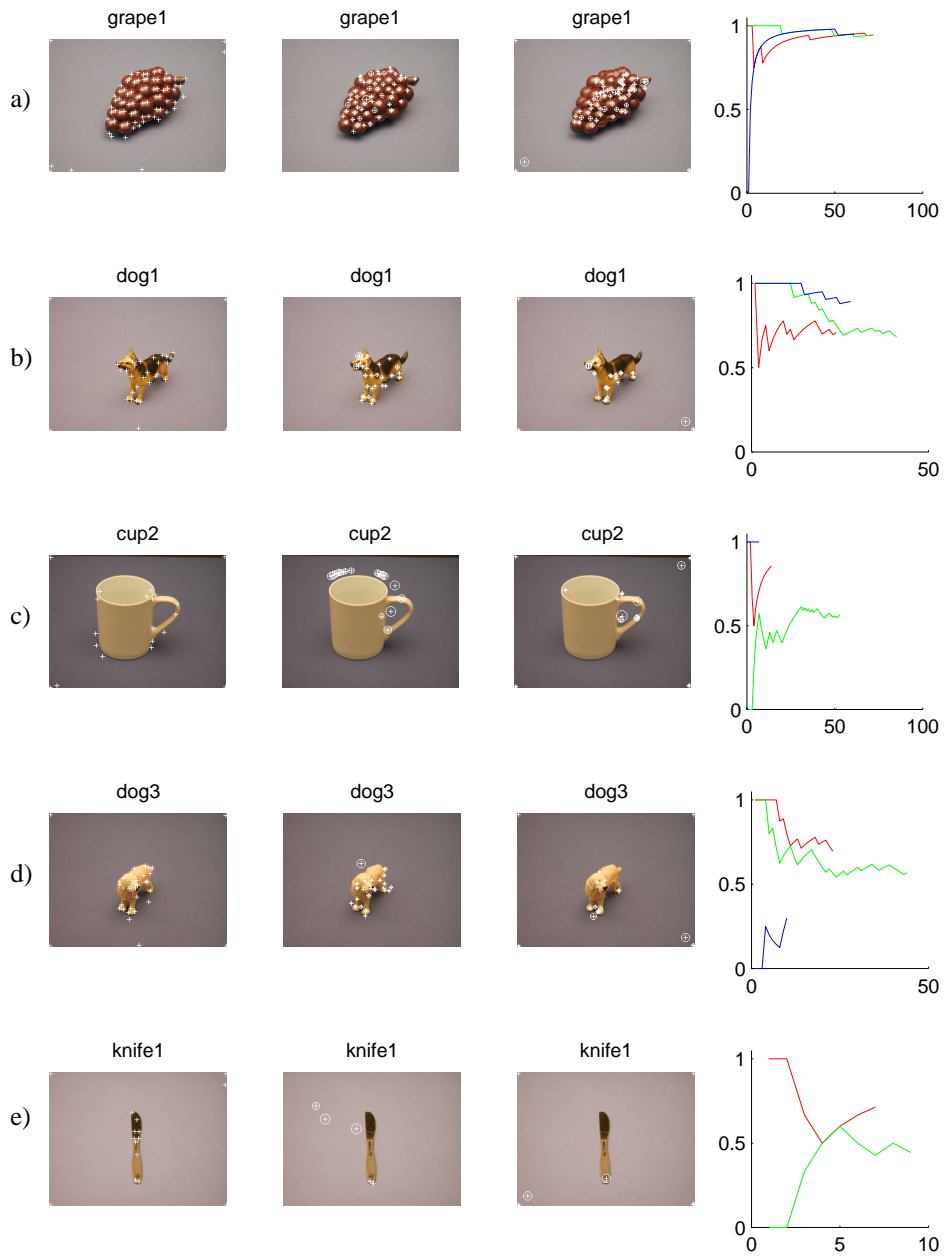
Figure 5: *Example images from the image database. Columns 1 to 3 show the points selected by the detectors Harris, Lindeberg, and Harris-Laplace respectively. The fourth column displays the probabilistic voting score after* n *votes (dark = Harris-Laplace, medium = Harris, light = Lindeberg)*