Scale-invariant shape features for recognition of object categories

Frédéric Jurie and Cordelia Schmid
GRAVIR, INRIA-CNRS, 655 avenue de l'Europe, Montbonnot 38330, France
{Frederic.Jurie, Cordelia.Schmid}@inrialpes.fr, http://lear.inrialpes.fr

Abstract

We introduce a new class of distinguished regions based on detecting the most salient convex local arrangements of contours in the image. The regions are used in a similar way to the local interest points extracted from gray-level images, but they capture shape rather than texture. Local convexity is characterized by measuring the extent to which the detected image contours support circle or arc-like local structures at each position and scale in the image. Our saliency measure combines two cost functions defined on the tangential edges near the circle: a tangential-gradient energy term, and an entropy term that ensures local support from a wide range of angular positions around the circle. The detected regions are invariant to scale changes and rotations, and robust against clutter, occlusions and spurious edge detections. Experimental results show very good performance for both shape matching and recognition of object categories.

1. Introduction

Local invariant features based on gray-level patches have proven very successful for matching and recognition of textured objects [14, 15, 20]. However, there are many objects for which texture is not a reliable recognition cue, but whose shape is highly characteristic. In particular, in category-level recognition [8], local shapes are often the most discriminant features shared by different instances of the category.

This paper introduces a new class of shape based salient regions and applies them to category-level object recognition. Our regions capture local shape convexities in scale-space. They are invariant to scale changes and rotations by construction, and illumination invariance is ensured by basing detection on image contours. Existing edge-based local descriptors [2, 4, 16] are based on local neighborhoods of points, whereas ours characterize the local shape near the boundary of a circle (*i.e.* within a thin annular region near the circle not the disk inside it).

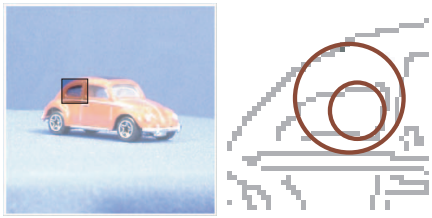

Figure 1: Two shape features detected at different scales.

To detect salient local shape convexities, we search over position and scale for salient circles — ones that receive locally maximal (over position and scale) support from the edges falling near their boundaries. Support is measured by combining two terms based on the edges near the circle: a classical edge-energy term that encourages strong, tangentially aligned edges, and a novel entropy term that ensures that the support comes from a broad range of angular positions around the circle, and not just from a few isolated points with unusually strong edges. Figure 1 shows a typical detection. Two regions were detected at different scales, each receiving support from several tangentially-aligned sections of local contour. Note that these two regions do not overlap: our local image descriptors are based on the distribution of contour points in a thin annulus near the circle boundary, not on the disk inside it. Apart from this, our descriptors are similar to shape contexts [2] and edge probes [4], which have previously been shown to be robust. As we will show, our descriptors allow object categories to be matched and recognized efficiently.

1.1. Related work

Computer vision research on edge and shape descriptions has a long history. In the 80's and 90's, approaches using alignment of edge points [9], global shape descriptors (Fourier transforms, skeletal shape, *etc*) [3, 5, 21], and geometric invariants [11, 18] were developed. These methods have significant limitations when used for real scenes. Alignment must search the space of all possible correspon-

dences between model and image edges, which becomes computationally prohibitive if the image and/or the model contains a large number of edges. Global descriptors require a good prior segmentation of the object and are not robust to occlusions. Geometric invariants are sensitive to missing or imprecise primitives.

More recently, several authors have proposed local shape descriptors that provide more robust recognition results. Selinger & Nelson [19] extract dominant image curves and use them to normalize the local image patch with respect to scale and rotation. The edges in the patch are then the description. Another semi-local approach is Belongie & Malik's shape contexts [2], which characterize the local image by histogramming its edges into radial-polar bins. Carmichael & Hebert [4] take a similar approach, characterizing each edge pixel by the local distribution of edges in its image neighborhood. Similarly, Mikolajczyk *et al* [16] measure the distribution of orientation information in a neighborhood chosen to be invariant to scale changes.

The approach developed in this paper is similar, but it has a different notion of locality, basing both detection and description near a circle rather than at all points within a solid patch. This makes it robust to events within the circle, *i.e.* it actually has *greater* locality than an equivalent 'point' detector running at the same spatial scale. For example, Laplacian interest points are often used to detect blob-like image structures, but they tend to fail when there is clutter within the blob (*e.g.* as for the large circle in fig. 1). Similarly, our approach tends to follow object contours, so it may be more robust to background clutter than boundary-based region descriptors such as shape context, for which half of the descriptor window often lies on the background. Furthermore, it does not rely on the extraction of long image contours, so it is significantly more robust than [19].

Organization: The paper is organized as follows. In section 2 we describe the extraction of our convex shape regions and compare them to gray-level based interest regions. Section 3 explains our shape region descriptors and gives some object matching results. Section 4 presents category recognition experiments.

2. Scale-invariant shape regions

This section describes how we extract our shape regions. An experimental comparison shows that our detector gives better results than gray-level based interest regions: it captures more of the shape information, and detection is more repeatable.

The key idea is to search across position and scale for salient local convexities of shape. Convex regions necessarily have good compactness, and they tend to be perceived as figure whereas concave ones are perceived as ground (Metzger's "law of the inside" [17]). Various classical measures

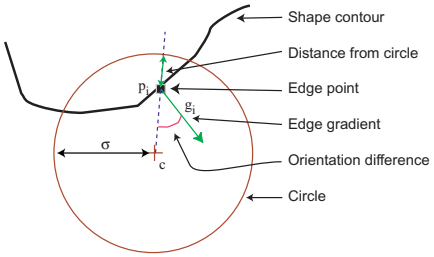

Figure 2: Quantities used for region extraction.

of convexity exist, such as the ratio between the area of the region and that of its convex hull, or the fraction of the region's boundary that lies on the convex hull's boundary [1]. Our measure is different and less reliant on having a complete boundary.

2.1. Region detection

Contour extraction. To detect image contours, we use the Canny-Deriche detector [7] at each level of our multi-scale representation, see figure 3 for an example of the contours extracted. At a given scale, the image is thus represented by N contour points $\mathbf{p}_i$, each with an attached image gradient vector $\mathbf{g}_i$, normalized to magnitude 1.

Contour point weights. For each position $\mathbf{c}$ and scale σ in our search space, we take the circle with this position and scale (*i.e.*, centre and radius), and search for all contour points $\mathbf{p}_i$ that lie near the circle. Our salience metrics are normalized weighted sums of local contributions over these contour points, with the weights set to reflect closeness to the circle and alignment with its local tangent. We measure closeness by the Gaussian of distance from the circle

$$w_i^d(\mathbf{c},\sigma) \equiv \exp\left(-\frac{(\|\mathbf{p}_i-\mathbf{c}\|-\sigma)^2}{2(s\sigma)^2}\right)$$

where s defines the locality of the detection. For small s only points close to the circle are taken into account. The value used in the experiments is $s = 0.2$.

Tangency is determined by the local image gradient's dot product with the local radial unit vector

$$w_i^o(\mathbf{c},\sigma) \equiv \left|\mathbf{g}_i \cdot \frac{\mathbf{p}_i-\mathbf{c}}{\|\mathbf{p}_i-\mathbf{c}\|}\right| = \|\mathbf{g}_i\| \cos\angle(\mathbf{g}_i, \mathbf{p}_i-\mathbf{c}).$$

The final weight for the point is then the product of the closeness and alignment weights

$$w_i(\mathbf{c},\sigma) \equiv w_i^d(\mathbf{c},\sigma) w_i^o(\mathbf{c},\sigma)$$

Saliency metric. Our saliency measure for circular regions is a product of two terms, the **tangent edge energy** which measures the extent to which the detected edges are strong and well-aligned with the circle

$$E(\mathbf{c}, \sigma) \equiv \sum_{i=1}^{N} w_i(\mathbf{c}, \sigma)^2$$

and the **contour orientation entropy** which measures the extent to which the circle has support from a broad distribution of points around its boundary (and not just from a few points on one side of it)

$$H(\mathbf{c}, \sigma) \equiv -\sum_{k=1}^{M} h(k, \mathbf{c}, \sigma) \log h(k, \mathbf{c}, \sigma)$$

where gradient orientation is quantized into M bins $k = 1, \ldots, M$ (32 in our experiments), and the contribution from each bin is

$$h(k, \mathbf{c}, \sigma) \equiv \frac{1}{\sum w_i(\mathbf{c}, \sigma)} \sum_{i=1}^{N} w_i(\mathbf{c}, \sigma) K(k - \tfrac{M}{2\pi} o_i)$$

Here, o_i is the angular orientation in radians of the contour gradient vector $\mathbf{g}_i$, and $K(x)$ is a smoothing kernel, here $K(x) \equiv \exp(-x^2/2)$.

Our final saliency metric for the circle is the product

$$C(\mathbf{c}, \sigma) \equiv H(\mathbf{c}, \sigma)\, E(\mathbf{c}, \sigma). \quad (1)$$

This captures the notion of convex saliency well, as E is large when there are strong edge points close to the circle and well aligned with its tangent, and H is large when the highly weighted edge points are well spread in orientation around the circle. It is maximal when the image contours correspond exactly to a complete circle with centre $\mathbf{c}$ and radius σ. There is no compelling theoretical reason for using the product to combine E and H in (1), but in practice this heuristic method of combination seems to work well.

Overall detector. To complete the detector, we calculate the circle saliency at all image positions and scales, find local maxima of saliency over position and scale, and define interest regions at each sufficiently strong local maximum of our saliency measure. Our scale space is built by filtering and down sampling the input image and extracting contours and calculating saliency at each scale, so our integration circles actually have a constant radius of 5 pixels (corresponding to circles of various sizes in the input image). For the experiments we used 30 scale levels spaced by a factor 1.1 in scale. We required detections to be the dominant local maximum over a suppression window of size 9×9 spatially and 3 in scale. Maxima were also discarded if their strength was less than 10% of the maximum strength observed in the image. Our current (unoptimised) detector takes about one minute to process a 256×256 image.

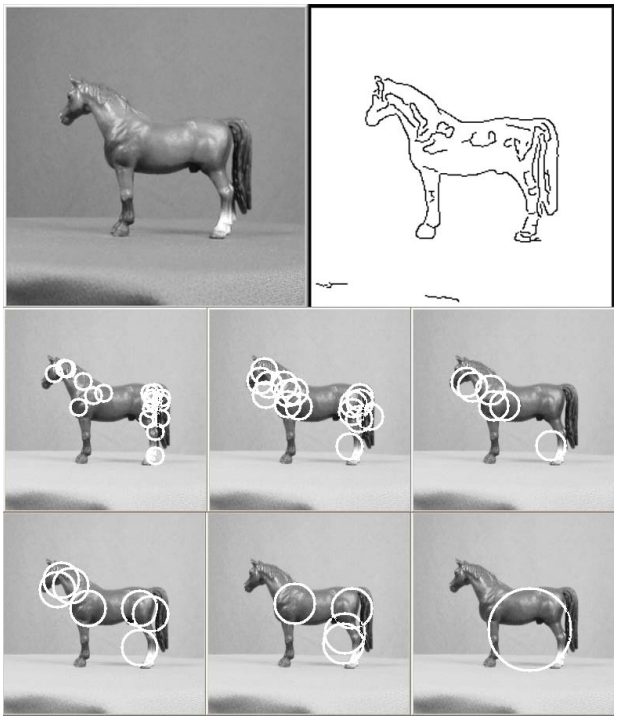

Figure 3: Top row: input image and extracted contours. Bottom rows: Regions detected at different scale levels.

2.2. Results

Figure 3 presents some typical detection results. The top row shows the image and the extracted contours. The remaining two rows show the regions extracted at scales (equivalent circle radii in the input image): $10, 15, 20, 22.5, 25$ and 50 pixels. Around 100 regions are detected in this image, but the figure only shows about 50 of them. Note that our shape features detect the most characteristic components of the shape at each scale.

Figure 4 shows that rescalings and background clutter have little effect on the detected features. Only the features at relatively large scales are shown here, because these are the most likely to be influenced by background clutter. For example, the region between the legs of the horse is still detected, despite the introduction of the grass. The results in section 4 below show that noise, clutter and occlusions also have little influence on recognition.

2.3. Comparison with interest region detectors

There are several available detectors for extracting scale-invariant interest regions in gray-level images. The ones that measure saliency in a way most similar to ours are the Laplacian detector [13] and the entropy detector [10]. The

Laplacian detector is designed to detect localized blobs in scale-space. As with our detector, it does detect perfect circles, but in real images the differences in performance are clearly visible — see figure 5. The two images represent two instances of the same object category (horses). The more similar a given detector's responses are on the two images, the more effective it is likely to be for category-level object recognition. Visual examination of the figure shows that the Laplacian (B) and Entropy (C) detectors tend to focus on surface markings, and hence give low repeatability, whereas our detector focuses more on shape and hence gives better repeatability. The Entropy detector seems to be a little better than the Laplacian one, but still much less consistent than ours.

Quantitative comparison bears these conclusions out: if we manually evaluate the percentage of similar detections, *i.e.* the percentage of regions detected at similar relative positions and scales in the two images, the score for the Laplacian is less than 10%, about 20% for the entropy detector and more than 60% with our detector.

3. Descriptors and matching

This section describes the circularly-supported image descriptors that we have designed for use with our circular salient region detector, and gives an example of image matching using these descriptors.

3.1. Descriptors

In order to match shape regions we need to characterize their distinctive content, while still remaining robust to small shape variations. Our descriptors are designed to capture the distribution of contour points in a thin annular neighbourhood of the region's defining circle. They are coarse bi-dimensional histograms representing the local density of contour pixels at given distances from the circle and angular positions around it. The bins represent 32 angular sectors and 4 radial rings spaced exponentially in radius, giving a final descriptor dimension of 128. Apart from the spatial reorganisation of the bins to characterize the edges near a circle rather than the edges in a closed region, our descriptors are similar to *shape contexts* [2] and *edge probes* [4], which have both proved to be robust descriptors for practical matching.

3.2. Matching example

To demonstrate that our detection and description can find and match corresponding components of similar but different shapes, we show two image pairs from the ETH-80 database [12] in figure 6. Circular shape features are extracted in each image and matched between images by

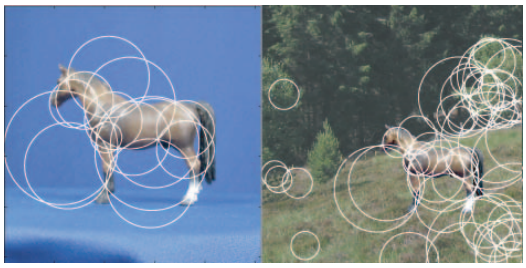

Figure 4: Detection of shape regions in the presence of background clutter. The right image contains background clutter, and the horse is about twice as large in the left image as in right one.

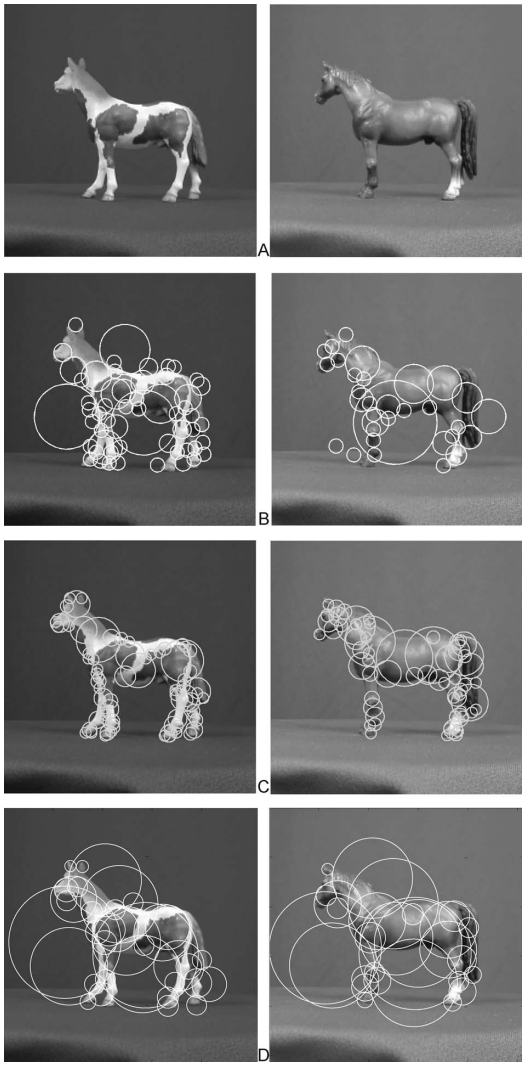

Figure 5: (A) Two different instances of the horse category. (B) Laplacian-based detection. (C) Entropy-based detection. (D) Our shape regions.

minimizing the χ^2 distance between the descriptor vectors. The 12 closest matches are displayed as circles in figure 6 and as the corresponding pairs of extracted patches in figure 7. The correspondence between the extracted patches is excellent.

4. Detection of object categories

In this section we apply our shape features to the detection of object categories. We first present the detection scheme, then we show results for three different object categories: cars, bikes and horses.

Our object models are trained with a few sample images. The experimental results show that this is sufficient, and confirm that our shape features capture the underlying object shape well. For each training image we extract the shape features and compute the descriptors. Object position and scale are known for the training images, which allows the relative position and scale of the extracted features to be fixed with respect to the model reference frame.

During object detection, we detect shape features in the test image, and match them to the database, *i.e* to the features extracted from the training images. Similarity between features is computed using the χ^2 distance, and all matches above a similarity threshold are kept. Many of these initial matches are incorrect owing to clutter. Each match then votes for the position and scale of the corresponding hypothesized object. We search this continuous voting scale-space for dense local clusters of hypotheses using Mean-Shift Mode estimation[6]. We use the recovered maxima as hypotheses, and verify each of these by aligning the model with the image and by counting the number of primitives that correspond. If this number is above a threshold, the object is declared to be present at the given position and scale.

For the car category we used 10 training images from the ETH-80 database[12], see figure 8. Figure 9 illustrates the different steps of the algorithm. The top row shows the image and the extracted contours. The second row shows the initial matches (left) and the matches that survive the verification step. The last row shows the gray-level patches corresponding to the selected features of the test image (left), and to their best matches from the training database (right).

Figure 10 shows similar results for several test images involving viewpoint changes (A), changes of object form (B), and noise and occlusion (C).

For the bicycle images, the training dataset consists of a single image, shown in the top row of figure 11. The remaining rows show test images with their detected (left) and matched (right) shape features.

The horse category is represented by five training images. The test set includes about 800 images: 400 with horses, and 400 including landscapes and people but no

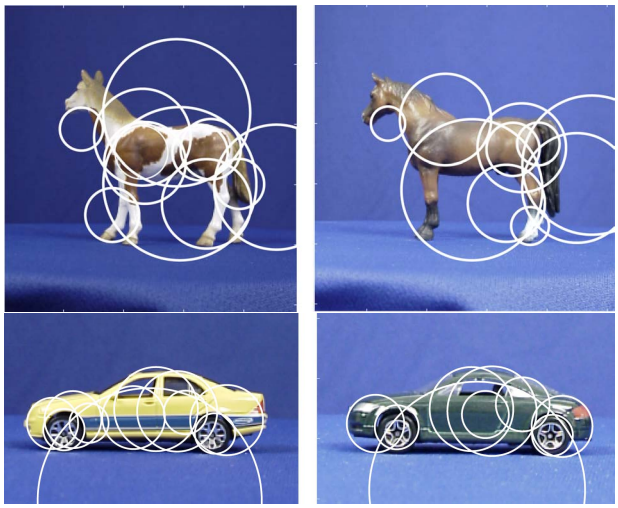

Figure 6: For each image pair (horse and car) the 12 matches with the best scores are displayed. Figure 7 shows the individual matching pairs of image patches.

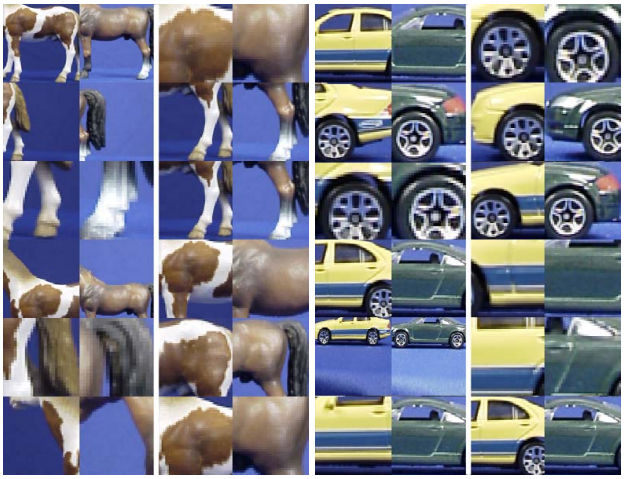

Figure 7: The individual region matches from figure 6. Note the excellent shape correspondence.

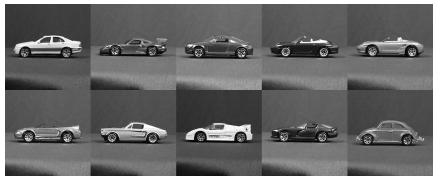

Figure 8: Car training images.

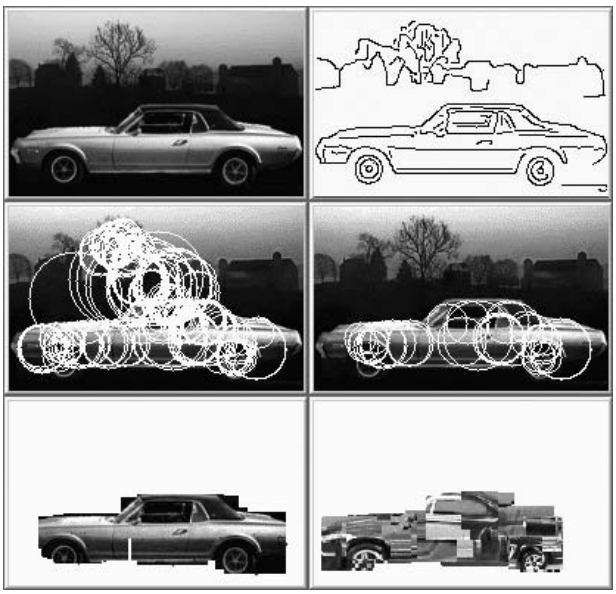

Figure 9: Car detection example. Top row: a test image and its contours. Second row: initial matches (left), matches after verification (right). Bottom row: assemblies of image patches corresponding to the test image features matched (left), and the training features that they were matched to (right).

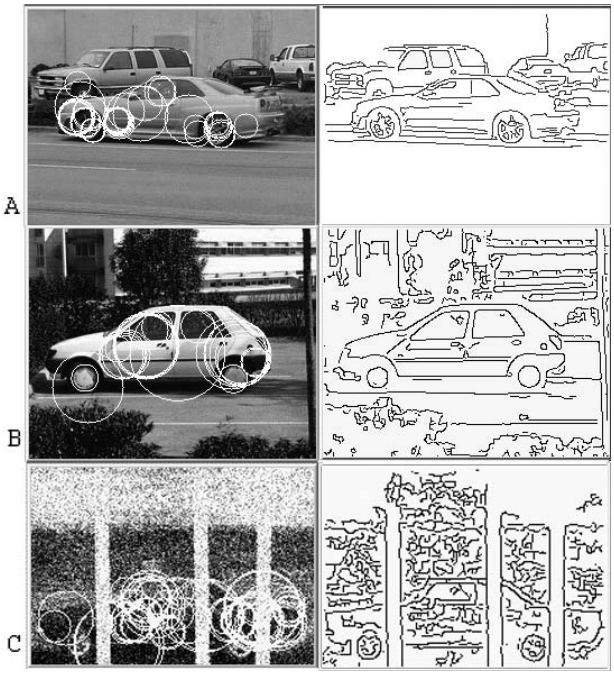

Figure 10: Detection results and the corresponding contour images. A) Change of viewpoint. B) Change of form. C) Noise and occlusion.

Figure 11: Bicycle detection. Top: Training image and the corresponding contours. Below: the detected features (left), and the matched features and detected objects (right), for two test images. Note that the bicycles are detected even under significant changes of scale factor.

% Precision	97.4	93.1	90.2	87.7	77.3	60.5
% Recall	23.2	46.7	63.4	70.4	82.5	96.4

Figure 12: Horse detection results. Top: images from the horse test set. Middle: precision and recall for several detection thresholds. Bottom: detections on a few test images.

horses. Figure 12 (top) shows some examples of this test set. In the middle row, the results are evaluated in terms of precision (the percent of detections that are correct) and recall (the percentage of correct images that are retrieved). The bottom row shows detections on a few test images.

For cars, there are about 2000 features in the database, as compared to about 15 for bicycles or 75 for horses. Around 100 to 1000 features are detected in the test images. The first matching step generates several hundred votes, but typically, less than 10 peaks in the transformation space have to be verified.

5. Conclusion and Discussion

In this paper we have introduced a new kind of shape feature based on annular regions, that has very good performance in the presence of occlusion and clutter. Our region detector is based on recent results in interest point detection, but adapts them to shape description. The results for category-level object detection are very promising.

Further research will include an extension to affine-invariant shape features. Affine invariance can easily be obtained by using an ellipse instead of a circle as the convex shape in the detection process. However, with affine transformations (up to rotation) the search space becomes 5-dimensional, so it will be more more efficient to first detect scale-invariant features, then to upgrade to affine invariance using a subsequent optimization step.

Our extracted regions are rotation invariant, but not our descriptors. However, preliminary experiments have shown that the descriptor can easily be represented relative to its dominant orientation (the dominant local edge orientation), which allows rotation invariance to be achieved. In addition, instead of thresholding edges and using the contour image, we could keep all edges and take into account the gradient magnitude.

Furthermore, our current detection strategy is straightforward and there are many avenues for improving it. We could for example construct a model based on the training images and determine which of its features are the most discriminative. We could also incorporate local neighborhood support into the matching process, and learn the uncertainty associated with the different features.

References

[1] K. Abe, C. Arcelli, T. Hisajima, and T. Ibaraki. Parts of planar shapes. *PR*, 29(10):1703–1711, October 1996.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(24):509–522, 2001.

[3] H. Blum. Biological shape and visual science. *Theoretical Biology*, 38:205–287, 1973.

[4] O. Carmichael and M. Hebert. Shape-based recognition of wiry objects. In *CVPR03*, pages II: 401–408, 2003.

[5] C. C. Chang, S. M. Hwang, and D. J. Buehrer. A shape recognition scheme based on relative distances of feature points from the centroid. *Pattern Recognition*, 24(11):1053–1063.

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5):603–619, May 2002.

[7] R. Deriche. Using canny's criteria to derive a recursively implemented optimal edge detector. *IJCV*, 1(2):167–187, 1987.

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR03*, pages II: 264–271, 2003.

[9] D.P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, November 1990.

[10] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 45(2):83–105, November 2001.

[11] Y. Lamdan, J.T. Schwartz, and H.J. Wolfson. Object recognition by affine invariant matching. In *CVPR88*, pages 335–344, 1988.

[12] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR03*, pages II: 409–415, 2003.

[13] T. Lindeberg. On scale selection for differential operators. In *SCIA93*, pages 857–866, 1993.

[14] D.G. Lowe. Object recognition from local scale-invariant features. In *Proc. IEEE International Conference on Computer vision*, pages 1150–1157, 1999.

[15] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV01*, pages I: 525–531, 2001.

[16] K. Mikolajczyk, A. Zisserman, and C. Schmid. Shape recognition with edge-based features. In *British Machine Vision Conference*, 2003.

[17] S. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, 1999.

[18] C.A. Rothwell, A. Zisserman, D.A. Forsyth, and J.L. Mundy. Canonical frames for planar object recognition. In *ECCV92*, pages 757–772, 1992.

[19] Andrea Selinger and Randal C. Nelson. A perceptual grouping hierarchy for appearance-based 3D object recognition. *Computer Vision and Image Understanding: CVIU*, 76(1):83–92, 1999.

[20] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affinely invariant neighbourhoods. *International Journal of Computer Vision*, to appear, 2003.

[21] C. T. Zahn and R. Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, 21, 1972.