An Approach to the Parameterization of Structure for Fast Categorization

Accepted at the International Journal of Computer Vision, 2009

Affiliation and mailing address:
Christoph Rasche
Institut für Psychologie
Justus-Liebig Universität
Otto-Behagel-str. 10, F1
35394 Giessen, Germany

fax: +49 641 99 26 119
phone: +49 641 99 26 106
email: rasche15@gmail.com

Abstract

A decomposition is described, which parameterizes the geometry and appearance of contours and regions of gray-scale images with the goal of fast categorization. To express the contour geometry, a contour is transformed into a local/global space, from which parameters are derived classifying its global geometry (arc, inflexion or alternating) and describing its local aspects (degree of curvature, edginess, symmetry). Regions are parameterized based on their symmetric axes, which are evolved with a wave-propagation process enabling to generate the distance map for fragmented contour images. The methodology is evaluated on three image sets, the Caltech 101 set and two sets drawn from the Corel collection. The performance nearly reaches the one of other categorization systems for unsupervised learning.

Keywords: contour description; curve partitioning; symmetric-axis transform; image classification; basic-level categorization

1 Introduction

Designing representations of human visual categories is difficult because category instances are structurally variable, e.g. the structures of chairs or living room scenes contain subtle geometrical differences, which are difficult to capture in a model. The variability persists at different levels ranging from the single contour to the entire configuration of features or parts (figure 2 in (Basri and Jacobs, 1997), p.452 in (Palmer, 1999), figure 4 in (Rasche, 2005), (Draper et al., 1996)). Design approaches have addressed this variability in various ways either implicitly or explicitly. At a lower level, the suggested representations range from 3D cylinders for parts (Marr, 1982), to template matching for contours - whereby the distance measure buffers the variability - (Shotton et al., 2005; Opelt et al., 2006; Fergus et al., 2004), to region information to circumvent contour variability (Kimia et al., 1995; Basri and Jacobs, 1997; Keselman and Dickinson, 2005; Felzenszwalb and Huttenlocher, 2005). At a higher level, flexible structural descriptions (e.g. (Zhu and Yuille, 1996; Nelson and Selinger, 1998; Fergus et al., 2007)) and deformable templates (Mori et al., 2005) have been proposed. Many of these systems operate on a small set of objects or scenes with small structural variability and certainly provide the optimal representations and recognition mechanisms for those. A first attempt to expand to a larger set of images with objects exhibiting a larger variability was made by Leibe and Schiele (Leibe and Schiele, 2003), who compared different classification methods. The most comprehensive approach is the one by Perona's group (Li et al., 2006), which shows an exceptionally high categorization rate for 101 photographed object categories; yet even those objects are either of limited geometrical variability or show very similar texture (Li et al., 2006). Still, there is need for methodology, which is suitable for categorization of images of arbitrary content, such as a texture, shape, object or scene, depicted in line-drawings, cartoons or photos. The human categorization process is exceptionally good and blazingly fast at this task: it assigns an image structure to one of 5000 basic-level categories within a duration of 100-200 milliseconds (Potter, 1976; Biederman, 1987; Thorpe et al., 1996), a process sometimes called *fast categorization* in psychophysics literature, e.g. (Joubert et al., 2008). Leibe and Schiele give an elaborate introduction to the nature of this assignment process and to some of the difficulties that one faces, when developing such category representations (Leibe and Schiele, 2003).

Description One viewpoint is that representations must be based on contour information, even though contours seem to appear partially incomplete in gray-scale images (see (Elder, 1999) for a recent summary of arguments). Following this viewpoint, the most concise formulation of what the representation should be, was given by Nelson and Selinger, who envisioned that an object representation is a loose assembly of contour fragments (Nelson and Selinger, 1998), or as they term it the *Cubist approach*. The fact that the human recognition process can be easily deceived by structural visual illusions (Gregory, 1997) is evidence that this type of representation is also used by the human visual system. Nelson and Selinger apply the approach to object detection, for which contours are represented as templates (see e.g. (Shotton et al., 2005; Opelt et al., 2006; Fergus et al., 2004) for variants). A contour template is however an unlikely candidate for fast categorization of arbitrary image content because of its low degree of abstraction. A more promising direction is therefore to parameterize contours. For instance, Berretti et al. represent contour protrusions (corner; high curvature) as a two-dimensional vector of orientation and degree of curvature (Berretti et al., 2000). They also use a loose representation to describe their objects and apply their methodology to collections of line drawings with limited variability. We think that an extension of this approach to gray-scale images has the greatest chance of succeeding at fast categorization, in particular the representation of contours by multiple dimensions to buffer structural variability. The advantage of a multi-dimensional representation is that it avoids an a-priori classification: if a structure is slightly deformed, it still can be compared to previous instantiations by a mere distance function in the respective multi-dimensional space, whereas if the representation is based on classified features, there exists the potential of a combinatorial explosion to deal with structural variability (Draper et al., 1996).

Should one decompose contours only? Basri and Jacobs made the argument to rather use region description to avoid contour variability (figure 2, (Basri and Jacobs, 1997)). Zhang and Lu also favor region-based methods over contour-based methods (Zhang and Lu, 2004), arguing that the entire region of a shape serves as a better descriptor than its contours. A region description can express a shape more compactly than its contours but does so only coarsely. But when moving to gray-scale images and tackling the issue of categorization, both types of information are necessary. For instance, regions of natural scenes are often highly variable and are thus little characteristic to this category; it is rather the irregularity and fuzziness of their contours, which make them so distinct from many other categories. Analogous, the silhouette contour of an animal or shape may be the only way to quickly categorize an image. After all, contours of gray-scale images appear so fragmented that transforming the structure in multiple ways may be the key for fast categorization and not the extraction of just a single descriptor.

Goal and Scope The goal of this study is to introduce methodology that decomposes any structure (texture, shape, object or scene) into elementary contour segments and elementary regions (areas), which then are parameterized. This parameterization could theoretically be carried out with near-infinite accuracy to ensure that all possible contours and regions can be distinguished. Such high accuracy were necessary if an identification task was followed, e.g. the discrimination of subtly different shapes. But for the purpose of basic-level categorization a limited number of parameters should suffice. How those parameters are then optimally employed for each category - e.g. rather statistical or as multi-dimensional vector -, required a systematic testing and a flexible learning scheme, which we have not implemented yet. We therefore chose to test the parameters in two extreme forms, a statistical (histogramming) and a 'precise' (vectorial) description. The evaluation shows that our system can already perform equally good as other categorization systems when those did not use human-supervised learning (Fergus et al., 2007; Oliva and Torralba, 2001).

The task of fast categorization should not be confused with object search (object detection or localization) as in (Shotton et al., 2008; Heitz et al., 2009) for instance. In object search, mostly part of the image is interpreted only. In the task of fast categorization in contrast, the entire image content is interpreted (see also (Oliva and Torralba, 2001) for arguments for scenes). For that reason we intentionally chose a low image resolution for our own image set (192 x 128 pixels, see 5.1), in order to force the search for useful representations. It would seem that this is a disadvantage, because at such a low resolution not even the fine-scale contours sometimes reveal the image content to the human observer. Still, humans have no difficulty to categorize images at low resolution, likely because they make also use of appearance information such as the luminance contrast along contours or the luminance statistics of a region. Such appearance information is also exploited in this study.

2 Contour description

We firstly survey previous contour approaches (subsection 2.1) in order to understand why a novel decomposition and description is required (subsection 2.2). The proposed decomposition starts with a labeling technique (see 2.3.1), which was loosely envisioned by Fischler and Bolles (see 2.2.2) for the purpose of curve partitioning. But the technique is extended by a distance measurement, which transforms the contour geometry into a so-called local/global space (see 2.3.2). This space allows to extract two kinds of structural information: the locations of high-curvature points as well as geometric parameters suitable for contour representation (subsection 2.3.4).

2.1 Previous contour approaches

A common method to segment and describe contours for shape representation is the curvature scale space (Asada and Brady, 1986). Asada and Brady low-pass filter contours with Gaussian functions at increasing scales to detect points of highest curvature at which the contours are broken up and then classified. The study was carried out on a limited set of shapes whose segments could be described by analytical functions. A similar scale space was proposed by Mokhtarian and Mackworth (Mokhtarian and Mackworth, 1986; Mokhtarian and Mackworth, 1992) which operated on a larger set of shapes and which provided various forms of shape invariances. These methods continue to be refined and has also been applied to image retrieval of shapes, e.g. (Dudek and Tsotsos, 1997; Zhong and Liao, 2007).

Parent and Zucker focus in particular on curves (Parent, 1989), providing algorithms which precisely determine curvature parameters; they apply their methodology to angiograms, satellite images and fingerprints. Similarly, Wang et al concentrate on convex boundaries applied to medical images (Wang et al., 2007).

The studies mentioned so far are optimized for identification of a selected set of images. However for categorization it requires less precise reconstruction but rather the capability to assign an image to a 'coarse' category: for instance to discriminate between a fingerprint and a giraffe, the contour description does not require high precision.

Other contour approaches which are similar to our approach are the following. Neural networks describe contours by a hierarchical integration of their local orientations in a pyramid space (e.g. (Serre et al., 2007; Hansen and Neumann, 2004; Amit and Mascaro, 2003; VanRullen and Thorpe, 2002; Rolls and Deco, 2002)). Yet such a description is essentially template matching in a more refined form.

Tu and Zhu propose a framework, in which images are segmented into a layer of regions and a layer of contours (Tu and Zhu, 2006). Contours are further classified as free curves, parallel curves and trees. Free curves are transformed into a sequence of oriented bars. Parallel curves, groupings of contours, and trees are represented as a Markov structure. Their study does not explicitly address the issue of contour matching.

The study of Fonseca et al is also a vector-based approach (Fonseca et al., 2006): they describe an entire shape with a single vector, whose dimensions represent trigonometric parameters, such as the area and the perimeter. Their description seems to work well with closed-contour line-drawings, but has not been extended to shapes or structures as obtained from gray-scale images.

Martin et al approach the detection of natural scene contours by developing an algorithm based on psychophysical studies (Martin et al., 2004) (see also (Yuille et al., 2004)). In those studies, humans segment images after categorization has been completed (after 100-200ms). But to what extent image segmentation is used for fast categorization is unclear and is difficult to elucidate with current psychophysical methods. Some even claim image segmentation is not necessary for fast categorization, e.g. (Oliva and Torralba, 2001; VanRullen and Thorpe, 2002).

The contour-description method by Felzenszwalb and Schwartz is probably most similar to our approach (Felzenszwalb and Schwartz, 2007). The method initially determines the geometry for three points of a contour (the two endpoints and the mid point) and recursively does so by halvening the contour into two segments (see (Gnther and Wong, 1990) for an origin of the concept). This generates a very distinct contour representation that can distinguish between a large set of shapes. The method even fulfills one of the desired categorization aspects that are listed in the next section (swiftness). But it can neither identify individual contour segments as our proposed method does, nor can the contour descriptions be as easily compared as in our approach (e.g. distance measurement in multi-dimensional space).

2.2 Categorization Approach

2.2.1 Differing Aspects

The approach to (fast) categorization requires methodology which differs from the above cited approaches by three aspects in particular:

a) Irregular contours: An essential part of categorization is the description of natural contours (see also (Martin et al., 2004)). But also non-rigid objects can possess irregular contour geometry which is very distinct. Consequently, it may not be ideal to "completely" or "perfectly" partition contours as it has been pursued in shape recognition studies (e.g. (Asada and Brady, 1986; Mokhtarian and Mackworth, 1986)). Rather one should attempt to include this irregularity (or 'wiggliness') as part of the partitioning process. Toward that direction there exists already work, which proposes alternate versions of the curvature scale space (Fischler and Bolles, 1983; Bengtsson and Eklundh, 1991).

b) Late contour classification: Many studies perform a contour classification after curve partitioning: e.g. Asada and Brady classify contours into corners and smooth join, which in turn are used to create compound changes (crank, end, bump or dent) (Asada and Brady, 1986); Bengtsson and Eklundh classify them into straight arcs, curved arcs with sign of curvature, corners (tangent discontinuities) and points delimiting arcs, especially inflexion points (Bengtsson and Eklundh, 1991). As already mentioned above, this classification may prohibit the development of representations capable of buffering variability. Instead one should attempt to express contour geometries as vectors to circumvent any early classification. In short, classification of contours and shapes should occur as late as possible in the feature extraction process - if at all.

c) Swiftness: The contour decomposition should operate as swiftly as possible. Curve partitioning and contour transformation should return detailed but not necessarily high-precision description. A transformation operating in a single pass (feedforward) is desired.

2.2.2 Challenges

The challenges of contour description and partitioning have been best formulated by Bengtsson and Eklundh (Bengtsson and Eklundh, 1991), p.85 bottom: To obtain a good contour description, "[...] one is faced with two goals of conflicting nature: First, there is need for finding a qualitative description of overall shape, hence some simplification and/or smoothing must take place. This goal is important for recognition and for finding global structure. Secondly, there is need for high precision detection of certain characteristics.[...]". In their method, Bengtsson and Eklundh still employ Gaussian filtering - as in the curvature scale space -, which leads to a coarsening of the contour at larger scales. But this is in direct conflict with their second goal. Instead, any coarsening should be avoided and only a later simplification should be undertaken to obtain geometric parameter values. Moving toward that goal, Fischler and Bolles's proposed to merely analyze the contour without altering it (Fischler and Bolles, 1983), p. 1016. They specifically propose an algorithm which "[...] labels each point on a curve as belonging to one of three categories: a) a point in a smooth interval, b) a critical point, or c) a point in a noisy interval. To make this choice the algorithm analyzes the deviations of the curve from a chord or 'stick' that is iteratively advanced along the curve (this will be done for a variety of lengths, which is analogous to analyzing the curve at different resolutions).[...]". We pursue their type of labeling, yet in an alternate form for two reasons: 1) their third label appears redundant because a noisy interval consists of a sequence of critical points when observed at a more global scale (larger chord size); 2) their specific goal was curve partitioning but the final goal needs to be contour description, for which - in addition to the detection of critical points - one also needs to filter for arcs (bows). This is elaborated next.

2.3 Decomposition

In a first approximation, a curve can be regarded as an alternating sequence of bows and inflexions: a wiggly (often natural) contour consists of an irregular alternation of bows and inflexions; an 'oscillating' contour consists of an even alternation; an arc or L feature consists of only a single bow (with no inflexion). To parameterize such a description, a bow/inflexion labeling technique is employed (similar to the above mentioned one), which is extended by a distance measurement, creating thus distance signatures (subsection 2.3.1). Creating these distance signatures for a range of chord (window) sizes, transforms the contour

geometry into what we call the *local-global space*, or *LG space* (subsection 2.3.2). The LG space provides two types of information: the location of potential partitioning points (2.3.3) and geometric parameters for contour description (2.3.4). To determine the global geometry, the LG space is simplified into a 'spectrum', the so-called *fraction function*, that captures the alternating behavior of a contour. The global description is centered around three types of geometries: arc, inflexion and alternating (wiggly). To describe the fine structure of a contour, a number of local parameters are defined, which express for instance whether the contour is either smoothly undulating or rather a zig-zag.

2.3.1 Labeling, Signatures, Block Function

Given an (open) curve with arc length variable v, a chord (or window or segment) of fixed length ω is selected and its endpoints connected by a straight line ℓ. The maximal deviation (or amplitude) a_{max} between the selected segment v_ω and the straight line ℓ is determined. If the segment lies primarily on one side of the straight line ℓ, the segment is labeled a bow and the amplitude a_{max} is assigned to a 'bowness' signature $\beta(v)$. If the segment lies on both sides, it contains a critical point (change of sign; transition) and the segment is labeled an inflexion, and the amplitude a_{max} is assigned to an inflexion signature $\tau(v)$. Iterating this labeling process through the entire contour creates the signatures $\beta(v)$ and $\tau(v)$. The signatures are set to zero for the boundaries where the chord can not be applied.

Three examples of simple signature functions are given. For a (perfect) arc, the bowness signature is a rectangular function (figure 1a), the inflexion signature is 0. The term *signature block* is now used to describe the range of neighboring values which are above 0. For an arc there exists one such signature block. For a contour made of a single inflexion, the bowness signature consists of two such signature blocks; the inflexion signature consists of only one signature block (figure 1b). If the arc is a L feature, the bowness signature block is a 'bump' (figure 1c). We later refer to a block of the respective signatures as simply the *bowness block* or the *inflexion block*.

Block parameters To obtain a parametric description the segment's geometry, that is outlined by a bowness block $\beta^\sqcap$, a few simple algebraic manipulations suffice. Its degree of circularity $\zeta^\sqcap$ is given by the integral:

$$\zeta^\sqcap = \int \beta^\sqcap. \tag{1}$$

To distinguish whether the block describes a L feature or arc, we define a parameter edginess $\epsilon^\sqcap$. It is determined by multiplying the derivate of $\beta^\sqcap$ by a normalized, ramp function f^r, whose width is equal to the block size (with center value equal 0):

$$\epsilon^\sqcap = \beta^{\sqcap\prime} f^r. \tag{2}$$

The edginess value is largest for a L feature, it is 0 for a perfect arc and negative for a flat bow. The block's symmetry $v^\sqcap$ is determined by integrating the absolute difference between the first block half $\beta^{\sqcap_1}$ and its reversed second block half $\beta^{\check{\sqcap}_2}$, which then is normalized ($l_\sqcap$=block size):

$$v^\sqcap = \frac{1}{2l_\sqcap} \int^{l_\sqcap/2} |\beta^{\sqcap_1} - \beta^{\check{\sqcap}_2}|. \tag{3}$$

A value of 0 means complete symmetry, an increasing value corresponds to increasing asymmetry.

The block parameters are later used to describe the local parameters of the contour geometry (see 2.3.4).

2.3.2 Local/global Space, Fraction Function

The above signatures are created for a range of window sizes, $\omega \in [\omega_{min}, l_c]$ (ω_{min} = minimum window size, l_c = total arc length of contour), generating what is now called the *local/global* space, or *LG* space:

$$\beta_\omega(v), \quad \tau_\omega(v), \quad \omega \in [\omega_{min}, l_c], \tag{4}$$

Figure 2 shows the LG space for a wiggly arc: at a local level (for small window sizes), the contour is noisy and the bowness and inflexion signature alternate; with increasing window size, the bowness blocks start to dominate.

To move toward an abstraction of the LG space and hence the contour's global geometry, the *fraction function* is generated, which relates the amount (or strength) ϕ of a label with the window level ω, forming so the bowness- and inflexion-fraction functions, respectively:

$$\phi^{\beta}(\omega), \quad \phi^{\tau}(\omega). \tag{5}$$

The course of the fraction functions differ for the three contour geometries. For an arc, ϕ^{β} increases with increasing window size, whereas ϕ^{τ} decreases (figure 2, right column, graph 'Fraction'); the rate of increase and decrease depends on the degree of smoothness (or 'wiggliness') of the arc. For a contour consisting of a single inflexion, the course of the bowness- and inflexion-fraction function is reversed (ϕ^{β} decreasing, ϕ^{τ} increasing; see supplementary information, figure 1). For an irregular (wiggly) contour, the bowness-fraction function describes a bump with its maximum located at a medium window size, whereas the maximum for the inflexion-fraction function occurs later (see supplementary information, figure 2). Thus, the fraction functions will be used to derive global parameters of the contour geometry (see 2.3.4).

2.3.3 Partitioning

Seeking to describe irregular contours also faces the difficulty to discriminate when they are accidental, e.g. when a contour arbitrarily spans several objects and is not characteristic for the category. For that purpose, some type of curve partitioning needs to take place. There are two essential criterions for curve partitioning.

a) Extremely high curvature: if a contour contains an 'end' - a turn of 180 degrees - it is partitioned at the point of highest curvature, because an end outlines an area, which is described more precisely by a region representation (see section 3). An exception to this rule should be when the contour is part of a perfect arc larger than 180 degrees. Exemplifying the partitioning rule, an ellipse, an oval or a U-turn is partitioned into its 2 elongated arcs. After application of this rule, any contour appears either as elongated in a coarse sense, or as smoothly circular.

Ends can be easily detected by analyzing the bowness signature from local to global: whenever its amplitude exceeds the length of a half circle with radius equal to half the window size ($\beta_{\omega}(v) > \omega\pi/2$), then the location of the maximum amplitude is selected as a potential point of partitioning. To recognize whether an end is part of a large smooth arc, the block's circularity $\zeta^{\sqcap}$ is used. If the circularity value exceeds a threshold, it is assumed to be part of a circular structure, such as a smooth arc. But the rule also selects any near-circular symmetric shape like square, pentagon, hexagon and so on. To further discriminate between the two cases it requires a clearer understanding of how to partition and represent such shapes. However, such cases are extremely rare in gray-scale images where contours often appear fragmented. It has therefore not been pursued further in this study.

b) Symmetry-asymmetry: if a contour contains a segment which is very symmetric, it may be part of an object and is therefore preferably partitioned or extracted. Symmetry is a very general characteristics and can describe the symmetry of individual signature blocks or of a segment described by a sequence of blocks. A partitioning following the symmetry criterion has not been pursued either, as it also requires a clearer idea of how to express symmetry at a more global level. The issue may lie in a similar vein as the partitioning and representation of symmetrical shapes.

2.3.4 Geometric Parameters

We attempt to express the geometries as a set of scalar parameter values, such that the geometries can be compared by a distance measure, e.g. the radial-basis function. We nominally distinguish between global and local parameters: the global parameters describe the types of contour geometries as outlined above (subsection 2.3.2); the local parameters rather describe the fine structure of a contour.

Global parameters One possibility to capture the range of global geometries is by counting the number of alternations of the bowness blocks. For an arc the value is just one, for a single inflexion it is two, for a sinusoid of 540 degrees it is three, and so one. Using a single parameter for the global description runs the risk that the local dimensions dominate the distance measure. To prevent this the global geometry is classified into three 'elementary' types: arc, inflexion and alternating. Although we warned about such a

classification earlier, we still allow for combinations of these geometries by choosing soft criteria for two types:

- **arc** (a): A contour is classified as an arc if its bowness-fraction function increases continuously. The parameter value is either 0 or 1.

- **transition** (t): A contour is classified as a single inflexion if the inflexion-fraction function increases at a global level. The parameter value corresponds to the maximum value of the inflexion-fraction function.

- **alternating** (w): A contour is classified as alternating, if it did not clearly classify as either arc or transition. To determine the alternating value, it requires the selection of a window size, that represents the 'oscillating' characteristic optimally. This window size is found by taking the ratio between the bowness- and inflexion-fraction function and selecting the window size, whose ratio value is closest to one. For that 'optimal' window size, the logarithm of the number of bowness blocks is used as the alternating value w (see window size=7 for the alternating contour shown in the supplementary material, figure 2).

A contour can have scalar values for more than one of the three parameters. For instance, a sickle shape - a 180-degree arc with a small straight line extension - is classified as transition but also as alternating due to the slight increase in the inflexion-fraction function caused by the extension (see supplementary material, figure 3).

Local parameters The local parameters are determined using the bowness block parameters (subsection 2.3.1). One parameter describes the overall curvature of the contour, which can also be regarded as a global parameter in case of an arc, but is a rather local aspect for the transition or alternating geometry. The symmetry measure expresses only local regularity.

- **curvature** (b): is defined as the maximum amplitude of the bowness space multiplied by the sum of values for the arc and the transition parameter, $b = \max_{v,\omega}(\beta_\omega(v))(a + t)$. If the global geometry has been classified as alternating then the curvature just corresponds to the maximum, $b = \max_{v,\omega}(\beta_\omega(v))$.

- **edginess** (e): is defined as the average $\langle\rangle$ of all edginess values for all k bowness blocks across window size, $e = \langle \epsilon^{\sqcap}_{k,\omega} \rangle$.

- **symmetry** (y): is defined as the maximum value of all bowness blocks across window size, $y = \max_{k,\omega} \upsilon^{\sqcap}_{k,\omega}$.

Alternate definitions are possible and several variants were tested. For the present categorization study, the variants did not change the results significantly (but may so if identification was tested).

Representational Capacity We now sketch how the parameters can express previously suggested elementary features and use in particular Asada and Brady's terminology ((Asada and Brady, 1986), see their figure 1):

1) A corner (L feature) and smooth arc can be described by a high value for the arc parameter and differing values for the edginess parameter (a L feature has a high value for e).

2) A smooth join and crank are expressed as a high value for the inflexion parameter (if the element appears isolated) and differing values for the edginess parameter. (In Asada and Brady's figure 1, the crank consists of a very sharp angle and would be partitioned in our approach).

3) A bump is represented as an above-0 value for the alternating parameter and an above-0 value for the edginess value.

Whereby cases one and two are geometrically exact, case 3 is not expressed as accurately anymore. If the contour is a natural contour, then it may be irrelevant, but if it is part of an object or scene, a more specific description may be desired. In order to be more specific, one had to include more parameters, or to pursue structural relations.

2.3.5 Appearance Description

The appearance dimensions are derived from a number of aspects:

- **contrast** (c_m, c_s): is the luminance difference along the contour. For a given contour, the average and the standard deviation of the contrast difference for all contour points is taken, resulting in dimensions c_m and c_s, respectively.

- **fuzziness** (f_m, f_s): expresses the degree of fuzziness along the contour. Contours of natural scenes show often a high degree of fuzziness, in particular the ones lying within a textured region. Analogous to the

contrast values, the average and standard deviation is determined for each contour, but which are taken from a preprocessed image. Preprocessing occurs with a blob filter selecting isolated pixels or a neighborhood of pixels whose contrast values 'pop-out'.

isolation (r): quantifies the amount of region around a contour, or put differently the degree of isolation within a structure. Contours or shapes can appear isolated, for example the circle of a sun (or moon) in a landscape scene, or the rectangle of a picture in a room scene. This degree of isolation is characteristic as well and must be part of a contour description.

The dimension r is determined from a so-called isolation map IM, whose generation is explained in the next section (specifically 3.1). For each contour the value of r is the average of the corresponding pixels taken from IM.

2.3.6 Summary

The derived geometric parameters are called arc a, transition t, alternating w, curvature b, edginess e, symmetry y. In addition to those, the parameters orientation (o) and contour length (l) are used (subsection 2.3.4). The appearance parameters are called isolation r, mean contrast c_m, contrast variability c_s, mean fuzziness f_m and fuzziness variability f_s (subsection 2.3.5). The parameters values are scalar and are used to span a 13-dimensional contour vector $\mathbf{c}$,

$$\mathbf{c}(o, l, a, t, w, b, e, y, r, c_m, c_s, f_m, f_s), \tag{6}$$

which is used for descriptor matching (subsection 5.3). More parameters could have been extracted, that describe for instance the 'skewness' or 'flatness' of an arc, but it is assumed that these suffice to perform the categorization of the chosen image sets (see 5.1).

3 Region Description

Previous Approaches The goal of region description is to represent relations between contours by measuring and representing distances between them. The common method to do this is the symmetric-axis transform (SAT) (Blum, 1973), which has already been used in many ways. Asada and Brady generate smoothed local symmetries (Asada and Brady, 1986), by taking the pair-wise distances between pixels of parallel contours, a straightforward method to generate the symmetric axis (sym-axis). In other studies, the SAT model aims at evolving sym-axes which represent the shape with high accuracy, applied for instance to the analysis and/or representation of medical image structures, e.g. (Ogniewicz and Kubler, 1995; Zhu and Yuille, 1996; Zhu, 1999; Niethammer et al., 2004). The group by Zucker classifies the symmetric-axes into 4 orders (Kimia et al., 1995; Siddiqi et al., 1999; Pelillo et al., 1999): 1st order is a U-shaped curve (protrusion; half an ellipse), 2nd order is a neck (two half-way fused circles), 3rd order is an oval (bend) and 4th order is a circle (seed). For their work, these classified sym-axes represent elementary structures and are used to describe various shapes.

Categorization Challenges But as pointed out above already, such a classification bears the potential of a combinatorial explosion and one should attempt to express such contour relations as vectors as well. Toward that goal we also employ the symmetric-axis transform, but in a more general form. We implemented the transform such that it can also be evolved for open contours - and not only for closed ones, which allows us to use any area in an image that is characteristic for a category. In addition, we also use another distance relation, which has already been introduced in the contour section (dimension r under subsection 2.3.5). Both distance relations can be obtained from a distance map. Distance maps can be generated in a variety of ways (Rosenfeld and Pfaltz, 1968; Fabbri et al., 2008), but we are not aware of an implementation that generates the map for arbitrarily fragmented contour images. To achieve that, we will use a neurally inspired wave-propagation process.

3.1 Distance Relations

Given a binary contour image, $CM(x, y)$, its contours are propagated using a wave-propagation process (Rasche, 2007), which is carried out in a single sweep, meaning there are no iterations necessary to obtain

the sym-axes (see supplementary information, figure 4a and b). From the distance map, one can derive two types of distance relations. One is the distance relation between neighboring contours, which is given by the crests (rims) in the landscape. Those were called symmetric axes by Blum (1967, 1973). They can be easily detected by convolving the distance map with a high-pass filter, followed by thresholding. The other type of relation is the degree of isolation of a contour, which is rather a global distance relation. It is obtained by convolving the distance map with a negative-peaked high-pass filter.

Extracting sym-axes map (SM) The distance map DM is convolved (*) with a high-pass filter F_{high} and thresholded by the function θ to obtain the symmetric map SM:

$$SM(x,y) = \theta(DM(x,y) * F_{high}(x,y)) \tag{7}$$

SM contains scalar values only at the location of the sym-axes (see supplementary information, figure 4b). The optimal filter input represents the conic shape generated by the contour of an isolated circle (a single sym-ax point).

Determining the isolation map (IM) DM is convolved with a high-pass filter with negative peak, F^{-}_{high}, at the location of contour pixels, $CM(x,y) = 1$. The convolution output is now called an isolation map, IM:

$$IM(x,y) = \begin{cases} DM(x,y) * F^{-}_{high}(x,y) & , CM(x,y) = 1 \\ 0 & , \text{else} \end{cases} \tag{8}$$

Contours that are isolated with reference to the entire structure return a high value; contours internal to a structure return a low value (see supplementary information figure 4c). In particular, high curvatures show high values, because they emit radially propagating waves, which represent the optimal input to the negative-peaked high-pass filter. For a contour, the "isolation" distance is described by the dimension r, which is defined as the average of its corresponding IM values. In contrast to the SM, the IM represents how isolated a contour is with respect to the entire surrounding region and not only with respect to its nearest-neighbor contour. The optimal filter input is an isolated, single contour point. For an isolated straight line, the endpoints in IM show the highest value, with decreasing values toward the line's center.

3.2 Parameterizing the Symmetric-Axis Signature

Geometry The symmetric map is partitioned at points of intersections. Because the sym-axes are temporally evolved, one can explicitly use the temporal dimension of the symmetric axis, that is the distance values s in dependence of arc length v, now called the symmetric signature (see supplementary information, figure 5). To describe the signature as thoroughly as possible, the following parameters are used: its two endpoint values, s_1 and s_2; its mean value s_m; the angle α taken from the signature's slope. If the signature contains a minimum or maximum between the endpoints then the location and value of that extremum is taken, s_{fx} and p_{fx} respectively. In case of absence of an extremum, s_2 serves as this extremum information. An elongation parameter e is determined, which is defined as the ratio between the total arc length of the sym-axis in the image plane and the mean distance s_m. Orientation (o) and curvature (b) of the signature in the image plane are two additional geometric parameters. The curvature was defined as the maximal distance between the sym-axis points in the image plane and the straight line connecting its ends.

Appearance The same appearance parameters are added as for the contours, but are based on the area spanned by the two contours. In addition to the mean and standard deviation, the range value for the intensity values is determined as well, c_m, c_s and c_r, respectively. The same parameters are also determined for fuzziness, taken from the preprocessed image.

Summarized we have the following 15-dimensional vector for an area $\mathbf{a}$:

$$\mathbf{a}(o, \alpha, e, s_m, s_1, s_2, s_{fx}, p_{fx}, b, c_r, c_m, c_s, f_r, f_m, f_s). \tag{9}$$

4 Implementation

Contour description Window sizes were generated in increments of $\sqrt{2}$. The smallest window sizes were $\omega = [5, 7, 9, 11, 13]$ (number of contour pixels) for scale $\sigma = 1$ and $\omega = [5, 9, 13]$ for larger scales. The signatures are normalized by their window lengths (a_{max}/ω) - this is a crude approximation but computationally cheap. For that reason, the threshold for detecting potential ends was set heuristically (value=1.2) as well as the threshold for detecting circularity (value=90). ϕ is determined as the fraction of signature values that are above 0 for each window size, e.g. $\phi^{\beta}(\omega) = [\int_v sgn(\beta_\omega(v))]/l_c$.

Two DOG filters are used to determine the fuzziness values: a 3x3 pixel with standard deviations of 0.5 and 1.0; a 5x5 pixel with standard deviations of 1.0 and 2.0. The output of both image convolutions is summed to a single image, from which f_m and f_s are determined for each contour. The remaining dimensions are determined without notable issues.

To find points of highest curvatures, the edginess values of the individual signature blocks are integrated across window sizes forming so an edginess signature, whose maxima denote those points.

Region description The contours are propagated using a thresholding mechanism inspired by our neural network implementation of a propagation process (see (Rasche, 2007) for details). The wave consists of a subthreshold propagation process and an active propagation process. The high-pass filter F_{high} to compute the symmetric map SM is a (fixed-size) DOG with standard deviations of 0.833 and 1.5. This crude filter approximation is made for reason of simplicity, because the use of the optimal, distance-dependent filter is computationally much more expensive.

The negative-peaked high-pass filter F^{-}_{high} to compute the isolation map IM is a DOG with standard deviations of 4.833 and 5.1667. The large filter size ensures that even the contours of small objects on a contour-free background are distinctively encoded, e.g. the contour of a bird in the sky.

Images, processing time Contours are detected employing the Canny algorithm (Canny, 1986), using the implementation of Matlab's image processing toolbox. Other edge detection algorithms could have been employed instead. The dimension values were normalized to a range of 0 to 1.

The average processing times for a Caltech image at scale $\sigma = 1$ and an average resolution of ca. 200x300 pixels using an Intel 2GHz was: 390ms for the Canny algorithm; 392 ms for the appearance information (without region description); 1270ms for the extraction of contour segments from the image; 2500ms for the generation of the LG spaces and the derived spectrum and parameter description; 3225ms for contour propagation, sym-axis extraction and parameterization. Summarized, the entire computation for an image is approximately 7.7 seconds (including inbetween saving of data files), but our computations are not as optimized as the C-supported Canny algorithm. For scale $\sigma = 5$, the average processing time is 3.4 seconds.

5 Evaluation

The decomposition output was tested using two extreme forms of representation. In one form, the parameters were used merely in a statistical sense, by creating histograms of the parameter values for all descriptors of an image (subsection 5.2). In the other form, the parameter values were used as dimensions of a vector (as summarized in equations 6 and 9; subsection 5.3). The optimal category representation is likely a mixture of those two extreme forms of representation and is individual to each category. But to develop such individual category representations it required systematic testing on a larger scale, which has not been implemented yet. The performance of the system is therefore tested with these two extreme forms of representation only. For both types of analyses, the radial-basis function was employed as distance measure; and a number of image searches were carried out to demonstrate that the requirement for 'graceful degradation' for categorization systems is fulfilled (Leibe and Schiele, 2003).

5.1 Image sets

In order to be faced with the largest amount of structural variability possible, the decomposition success is evaluated on the Corel and Caltech 101 collection. The images of the Caltech collection contain primarily

single objects with either relatively clear silhouette and limited geometric variability or with very similar texture across instances (Li et al., 2006). The categories can therefore be regarded as subordinate categories. The images of the Caltech collection were downsampled to a size of approximately 300x300 pixels, most of which are already at that size.

To test also complex scenes - containing multiple objects and 'smeared' object contours -, the Corel collection was used. Two sets of images are drawn from it. One set is the Urban&Natural set as chosen by Oliva and Torralba, who selected images from 8 super-ordinate categories (Oliva and Torralba, 2001) (mountain scene, forest scene, street scene, highway scene,...; resolution of 256x256 pixels). The other set is chosen by us and is created as follows. By design, the images of the Corel collection are organized into 600 classes (100 images per class). Approximately 360 classes correspond to a human basic-level or subordinate category (Rosch et al., 1976; Oliva and Torralba, 2001). The classes are of all types of structure (textures, isolated objects, complex scenes) and this image set shows therefore the largest variability. The image classes were pooled into 112 basic-level categories. Examples of basic-levels are (in decreasing proportion): wild animals (27, 4.5%), patterns (25, 4.2%), sports (25, 4.2%), flowers (17, 2.8%), aircrafts (16, 2.7%), models (13, 2.2%), birds (11, 1.8%), water animals (10, 1.7%), cars (9, 1.5%), canyons (7, 1.2%), different cultures (7, 1.2%), mountain sceneries (7, 1.2%), ships (7, 1.2%). Such a categorization is sometimes ambivalent because many scenes can be assigned or interpreted in various ways as humans possess different *entry-level* categories (Jolicoeur et al., 1984), also called perception subjectivity. To avoid a strong perception bias, the classes were categorized by two persons (the author and a research assistant). We later refer to this simply as the Corel set. For this image set we chose a resolution of only 192x128 pixels.

The subsample size was 10 images per category for the Caltech and Urban&Natural set (1010 images per entire subsample) and 10 percent images per category for the Corel set (typically 10 or 20 images; 3570 images per entire subsample), for both the learning and testing procedure. The images for a subsample were selected randomly and categorization performance was tested with 3 different subsamples using cross validation.

5.2 Histogramming

For the two descriptors for a given image (contours and areas), a 10-bin histogram for each dimension is constructed. The histograms are then concatenated to form a 280-dimensional image vector. The image vectors for one category subsample were averaged. The performance for correct categorization was between 9.8 and 12.3 percent for all 4 scales for the Caltech and Corel set (figure 4, labeled 'F'). Omitting the distracter category ('Google' images in Caltech collection) decreased the performance by 0.3 percent only. For the Urban&Natural set the performance for correct categorization was ca. 40 percent for scale 3.

To estimate the contribution of the individual descriptors and its dimensions, the performance was also determined when only a subset of dimensions was used. When using only the contour parameters (130-dimensional image vector; labeled 'c'), the performance decreased to a value between 6 and 9 percent. For the geometrical parameters of the contour descriptor (90-dimensional vector, 'c-geo'), the performance remained approximately the same for the Caltech collection, but rather dropped for the Corel set down to 4 to 6 percent, which exposes the characteristic, that the Caltech set contains objects with limited geometric variability. When only the appearance parameters were employed ('c-app'), the performance decreased slightly for the Caltech set but remained about the same for the Corel set. This reveals that the strongest cue for the Caltech categorization performance is the set of geometric parameters, but for the Corel performance it is the set of appearance dimensions. A similar performance pattern can be observed for the area descriptors ('a', 'a-geo', 'a-app'). The complete area vector yielded a higher performance than the complete contour descriptor for both image sets. The performance for the appearance dimensions of the Corel set is marginally larger than the one for the complete descriptor and is even nearly as high as the performance for full dimensionality. Clearly, the representative power of the individual or group of dimensions do not add linearly in the histogramming approach.

To estimate the significance of individual dimensions, the categorization performance was tested when single dimensions were knocked out (270-dimensional vector, figure 5). The performance decreased only slightly (ca. 0.3 percent), which demonstrates that none of the dimensions is crucially more significant than any other one.

A large number of histogramming variants were tested, such as two-dimensional histograms - pairings of

dimensions - with up to 1000 dimensions in total, as well as different bin sizes. Various learning schemes were explored as well. But substantially higher performance was not achieved. It seems that a categorization performance of ca. 12 percent is a robust result but it also appears to be an upper limit.

Figures 6 and 7 in the supplementary material show image searches for the entire Corel collection (best and worst sortings shown).

5.3 Descriptor Matching

The purpose of this part of the evaluation is to explore how specific single vectors or group of vectors can be. In a learning phase, descriptors were searched, that are characteristic to a category, so called *category-specific descriptors*. They can be regarded as a precursor of a Cubist representation. The category-specific descriptors were obtained by sorting images using individual descriptors. In a testing phase, the list of collected category-specific descriptors were matched against the descriptors of individual images of another subsample to determine the strength of category selection. We also tested categorization performance using those category-specific descriptors, but we did not achieve substantially higher performance than with mere histogramming (previous subsection). For that reason we resorted to this type of image search.

In the learning phase, each descriptor (e.g. the contour vector of a L-feature of a chair) was compared to all other descriptors of the remaining images in the subsample and the distances sorted by decreasing similarity. The category-specificity of a descriptor was defined as the percentage of images belonging to the same category (chair) for the first 100 images of the corresponding first 100 similar descriptors. Only descriptors with a minimum specificity of 2 percent were kept, called the category-specific contours (figures 6 and 7; see figures 9-14 in supplementary material for all scales and both sets; for sym-axes see figures 15-20). The category specificity could reach up to tens of percent and was 3.5 to 8 percent in average: for the Caltech collection the average was 7.5 percent for the contour and area descriptors; for the Corel set the average percentage was lower by ca 1.0 for each descriptor. Differences across scales were small (ca. 1 percent). That the representative power of single descriptors is not confined to the selected categories is demonstrated with image sorting using the entire Corel collection (60000 images, figure 8; see figure 8 in supplementary material for an example for sym-axes).

In the testing phase, the descriptors $\mathbf{v}_j$ for each image, were matched against the collected category-specific descriptors $\mathbf{v}_i$ of each category, resulting in a distance matrix D_{ij}. The shortest distance for each collected descriptor was selected $\mathbf{d}_i = \max_j D_{ij}$. This distance vector reflects the optimal match between a selected image and the Cubist category representation. The distance vector $\mathbf{d}_i$ was sorted and the first 2, 5 and 10 differences summed ($\delta_2, \delta_5, \delta_{10}$ respectively), followed by determining the category-specificity for each δ as before. A systematic search was carried out for the maximally performing δ value, separately for contour and area descriptors for 3 different scales ($\sigma = 2, 3, 5$). The maximal value was in average 18 percent for the Corel set and 24 percent for the Caltech collection, demonstrating the high distinctness of the vectors. The performance difference for the two image sets can again be explained by the differing degree of structural variability of categories within the two image sets.

5.4 Summary

The performance for correct categorization was 10-12 percent for the Caltech and Corel set and ca. 40 percent for the Urban&Natural set. This seems low as compared to other categorization attempts (Oliva and Torralba, 2001; Fergus et al., 2007), but it should be emphasized that the achieved categorization percentages of those systems were obtained with help of human-supervised learning, that is individual features or descriptors were given as category-specific 'clues'. Thus, the performance of our system is well comparable to the other systems and is relatively high given that only histogramming matching was used.

As noted, we had also tried to perform a categorization task using the individual descriptors, but we were not capable yet to obtain a higher categorization performance. Instead, we demonstrated the specificity of the vectors by a search performance (subsection 5.3), which again turned out to be relatively high, given that the maximum selection took place only for performances for one descriptor type for one scale.

A number of robustness tests were carried out, such as a different distance function (Euclidean), or different definitions for some of the dimensions. All these variations did not alter overall performance

significantly, suggesting that the decomposition is generally applicable and is not biased toward a specific image set.

6 Discussion

6.1 Further Comparison to Other Approaches

Categorization is sometimes understood as part of a generic object recognition process (see Keselman et al. (Keselman and Dickinson, 2005) for a concise history of object recognition trends). But the human visual system solves different recognition tasks by different processes. Fast categorization is only carried out for *canonical* views (Palmer et al., 1981), which are structures seen from familiar viewpoints - as depicted in the Corel or Caltech collection for instance. And given the fact that the process can be easily deceived by visual illusions, one can regard fast categorization as a rather superficial judgment of its image content - and *not* a comprehensive understanding. In scene perception research there exist the term 'gist' for this superficial judgment (see (Oliva and Torralba, 2001)). For *non-canonical* viewpoints in contrast, the human visual system requires more time to categorize the object, a *late* categorization in some sense (Palmer et al., 1981). The additional time may be only tens of milliseconds, hardly noticeable to humans, but has already triggered a host of mechanisms starting to mentally manipulate the decomposed structure. Such manipulations may already be simulated in object recognition studies performing viewpoint independent recognition (e.g. (Brooks, 1981; Lowe, 1985)). A computer vision system performing both of these recognition tasks - and other visual recognition tasks - may be a futuristic goal. However, a system performing fast categorization only is not. If the appropriate loose (Cubist) vector representation can be found for those category levels, then the assignment of an image to one of the 5000 basic-level categories within a duration of 200ms is a feasible goal.

One may divide present approaches to image classification or image understanding into two opposite sides regarding their degree of structural reconstruction. One side pursues an exact reconstruction of the image, starting for instance with image segmentation and continuing with grouping operations (Marr, 1982; Witkin and Tenenbaum, 1983; Malik et al., 2001; Elder et al., 2003; Tu and Zhu, 2006). The aim of such approaches is to systematically extract scene information, which eventually leads to categorization, but actual transformations of structure have been pursued to a limited extent only (see (Sudderth et al., 2008) for image transformations for object detection). The other side attempts to avoid any elaborate reconstruction by preprocessing the image with 'simple' features or single transformations, whose output is then classified or matched (Oliva and Torralba, 2001; Renninger and Malik, 2004; Mori et al., 2005). The former approach may be hampered by the pursuit of perfect reconstruction, the latter by too little reconstruction. The present approach aims in between and attempts to rigorously transform structure without placing emphasis on perfect reconstruction. It follows the viewpoint that a multitude of mechanisms is required to recognition (Minsky, 1975), whereby the challenge is to find the precise mosaic of mechanisms, which transforms a structure into a highly distinct description within a multi-dimensional space allowing so for fast assignment to a category representation. The idea that the human visual system may use a multi-dimensional space for representation has been suggested by Mel (Mel, 1997).

The decomposition output is suitable for visual search, i.e. the selection of a region of interest, because structure is expressed as a list of vectors. The list of vectors can now be searched for variances, which represent potentially interesting points (or regions). Indeed, our decomposition model can explain all variances discovered in seminal studies of human visual search (Noton, 1971; Treisman and Gormican, 1988). Others have already implemented architectures to mimic such search behavior (Itti et al., 1998; Privitera and Stark, 2000; Rajashekar et al., 2008), however those models merely extract straight contour orientations and can therefore explain only a small number of those findings. In contrast, the presented model explains a much larger number of findings such as the computation of precise contour curvature, contour angle and aperture of an arc ((Treisman and Gormican, 1988), figure 5, 10 and 11 respectively).

Some of the dimensions developed here certainly have a relation to the ones developed in texture studies (Ravishankar Rao, 1993; Haralick, 1979; Tamura et al., 1978; Amadasun and King, 1989). For instance, the region dimension seems analogous to the coarseness property. In many of those studies, the creation of such properties is based on or is related to human judgment about the cognitive qualities of texture properties.

6.2 System Performance

The performance of our system reaches the one of other fast categorization systems when those did not use any type of human-supervised learning cues (see 5.4). However, our system was not particularly tuned to any image set yet; and the system operates on any image type (texture, shape, object and scene) which is particularly demonstrated with the relatively high search percentage for the Caltech set (see 5.3). But most importantly, our method allows for an understanding of image parts (components, regions), because the local/global space and the area segments are detailed representation (see also previous subsection). While in the other category systems (Fergus et al., 2007; Oliva and Torralba, 2001) a new preprocessing of the image had to take place in order to understand parts of it (see (Torralba et al., 2006) for instance), in our approach this is not necessary as a structural decomposition is the center piece of the methodology.

Although image search was not the focus of the study (figure 8; figures 6-8 in supplementary material), it makes it worthwhile relating to studies pursuing such tasks, e.g. (Heidemann, 2005; Wang et al., 2001; Mojsilovic et al., 2004; Vogel and Schiele, 2007). These studies use traditional techniques such as image segmentation, template matching and interest points. Some perform on large collections, searching for any type of image (Heidemann, 2005; Wang et al., 2001; Novak et al., 2008); others describe the content of a smaller number of images, but with a specialization in subordinate categories (Vogel and Schiele, 2007). Common to all these approaches is that their success depends to a large extent on the use of color information. We think that our image-search results compare well to theirs (e.g. figure 8 (Heidemann, 2005); figures 12-14 (Mojsilovic et al., 2004)), yet our results are obtained without the use of any color information. An important constraint of such an applied image search is that it has to occur fast. The present decomposition may require more computation time than the techniques in the other studies, but the image size in this study was only 192 x 128 pixels, less than half the size of the images in the other studies.

6.3 Outlook

A continuation of the present approach needs to address the following issues:

1) Combination of descriptors: A scheme for combining different descriptors needs to be developed, which searches for category-specific combinations of several descriptors, across different contours and regions, and also across different spatial scales. Studies using (unparameterized) contour segments for object description and search may serve as a source of inspiration for such schemes (Shotton et al., 2005; Opelt et al., 2006).

2) Partitioning: Although the display of collected category-specific descriptors suggests that the present partitioning rules seem to suffice (figure 6), those collected descriptors were probably the ones which were 'naturally' partitioned by a large contrast. However, the key to proper partitioning may be a (global) context analysis, that is a comparison of the LG spaces of all image contours, before specific partitioning points are chosen.

3) More abstraction: So far only contours and the region between two contours were abstracted. An attempt should be made to abstract also complex regions, for instance as outlined by the intersecting sym-axes (which were partitioned), as well as vertex and T-feature features (Lowe, 1985). The latter are rare but can be crucial for the description of some objects. The traditional method to detect them is to determine distances between contours, but they can also be found exploiting the methodology presented here, for instance by detecting proximal starting points of sym-axes (s_1).

4) Grouping: The SAT can be regarded as a local grouping mechanism, but grouping should also take place on a more global level. One possibility is to exploit the spatial scale: areas of a coarser scale often encompass the texture on a finer scale and an abstraction for the texture could be developed, e.g. histogramming. But this type of scale-based grouping may not be sufficient for a reason analogous to the analysis of contour geometry: the LG space was developed in order to circumvent the loss of structural information through smoothing. Similarly, one should explore global-grouping mechanisms applied to the same spatial scale. As the contours and areas are expressed as vectors, such grouping could be performed by mere vector analysis.

5) Appearance description: the choice of appearance dimensions is rather simple (luminance and fuzziness dimensions; c_m, f_m,...), but texture perception studies have shown that the detailed distribution of luminance values seems to be a strong determinant for proper texture identification (Dror et al., 2004; Motoyoshi et al., 2007). A more thorough parameterization of the luminance distribution should therefore be tested.

Summarizing, we propose that a many-to-many feature matching concept should be an essential part of fast categorization (see also (Demirci et al., 2006)), but this is not to argue for a specific representation. The present evaluation has shown that statistical (histogramming) and structural (descriptors) information are both very powerful for categorization. Given the presented decomposition, segmentation and grouping operations may be carried out in a novel light (Malik et al., 2001; Tu and Zhu, 2006; Elder et al., 2003), yet with the caveat not to pursue an exact reconstruction. Furthermore, the challenge of fast categorization may not be clearly separable from the task of retrieving a frame, which contains the information of spatial relations amongst objects or scene parts (Biederman et al., 1982). Thus, mechanisms such as scale selection (Berengolts and Lindenbaum, 2006), saliency detection (Kadir and Brady, 2001) or sequential pattern recognition (Fu, 1968), must be taken into account for a solution to fast categorization. Even template matching must be considered; given the enormous memory capacity of the visual system (Standing et al., 1970; Brady et al., 2008), a 'loose' vector-template with multiple descriptors may also be part of the solution. Such a representation may not be much smaller in size than the image per se, but the vector representation can deal with many little structural subtleties.

Acknowledgment

The study is supported by the Gaze-Based Communication Project (European Commission within the Information Society Technologies, contract no. IST-C-033816). The author wishes to thank Nadine Hartig for help with categorization, and Karl Gegenfurtner for lab support.

References

Amadasun, M. and King, R. (1989). Textural features corresponding to textural properties. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:1264–1274.

Amit, Y. and Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vision Research*, 43(19):2073–2088.

Asada, H. and Brady, M. (1986). The curvature primal sketch. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:2–14.

Basri, R. and Jacobs, D. (1997). Recognition using region correspondences. *International Journal Of Computer Vision*, 25(2):145–166.

Bengtsson, A. and Eklundh, J.-O. (1991). Shape representation by multiscale contour approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:85–93.

Berengolts, A. and Lindenbaum, M. (2006). On the distribution of saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1973–1990.

Berretti, S., Del Bimbo, A., and Pala, P. (2000). Retrieval by shape similarity with perceptual distance and effective indexing. *IEEE Transactions on Multimedia*, 2:225–239.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychological Review*, 94:115–45.

Biederman, I., Mezzanotte, R., and Rabinowitz, J. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14:143–77.

Blum, H. (1973). Biological shape and visual science .1. *Journal Of Theoretical Biology*, 38(2):205–287.

Brady, T. F., Konkle, T., Alvarez, G. A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *PNAS Proceedings of the National Academy of Sciences*, 105(38):14325–14329.

Brooks, R. (1981). Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence*, 17:285–348.

Canny, J. (1986). A computational approach to edge-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

Demirci, M., Shokoufandeh, A., Keselman, Y., Bretzner, L., and Dickinson, S. (2006). Object recognition as many-to-many feature matching. *International Journal Of Computer Vision*, 69(2):203–222.

Draper, B., Hanson, A., and Riseman, E. (1996). Knowledge-directed vision: Control, learning, and integration. *Proceedings of the IEEE*, 84(11):1625–1637.

Dror, R., Willsky, A. S., and Adelson, E. H. (2004). Statistical characterization of real-world illumination. *Journal Of Vision*, 4(9):821–837.

Dudek, G. and Tsotsos, J. (1997). Shape representation and recognition from multiscale curvature. *Computer Vision and Image Understanding*, 68:170–189.

Elder, J. (1999). Are edges incomplete? *International Journal Of Computer Vision*, 34(2-3):97–122.

Elder, J., Krupnik, A., and Johnston, L. (2003). Contour grouping with prior models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):661–674.

Fabbri, R., Da F. Costa, L., Torelli, J. C., and Bruno, O. M. (2008). 2d euclidean distance transform algorithms: A comparative survey. *ACM Computing Surveys*, 40(1):Article 2, 2:1–2:44.

Felzenszwalb, P. and Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal Of Computer Vision*, 61(1):55–79.

Felzenszwalb, P. F. and Schwartz, J. D. (2007). Hierarchical matching of deformable shapes. *IEEE Conference on Computer Vision and Pattern Recognition, 17-22 June 2007, Minneapolis, USA*, pages 1–8.

Fergus, R., Perona, P., and Zisserman, A. (2004). A visual category filter for google images. *European Confernence on Computer Vision 2004, PT1*, 3021:242–256.

Fergus, R., Perona, P., and Zisserman, A. (2007). Weakly supervised scale-invariant learning of models for visual recognition. *International Journal Of Computer Vision*, 71(3):273–303.

Fischler, M. and Bolles, R. (1983). Perceptual organization and the curve partitioning problem. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence, Volume 2*.

Fonseca, M., Ferreira, A., and Jorge, J. (2006). Generic shape classification for retrieval. In *Graphics Recognition. Ten Years Review and Future Perspectives*. Springer Verlag Berlin/Heidelberg.

Fu, K. (1968). *Sequential methods in pattern recognition and machine learning*. Academic Press, London.

Gnther, O. and Wong, E. (1990). The arc tree: An approximation scheme to represent arbitrary curved shapes. *Computer Vision, Graphics and Image Processing*, 51:313–337.

Gregory, R. (1997). Knowledge in perception and illusion. *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.*, 352(1358):1121–1127.

Hansen, T. and Neumann, H. (2004). Neural mechanisms for the robust representation of junctions. *Neural Computation*, 16(5):1013–1037.

Haralick, R. M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67:786–804.

Heidemann, G. (2005). Unsupervised image categorization. *Image And Vision Computing*, 23(10):861–876.

Heitz, G., Elidan, G., Packer, B., and Koller, D. (2009). Shape-based object localization for descriptive classification. *International Journal of Computer Vision*, 84:40–62.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 20(11):1254–1259.

Jolicoeur, P., Gluck, M., and Kosslyn, M. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, 16:243–275.

Joubert, O. R., Rousselet, G. A., Fize, D., and Fabre-Thorpe, M. (2008). Processing scene context: fast categorization and object interference. *Vision research*, 47(26):3286–3297.

Kadir, T. and Brady, M. (2001). Saliency, scale and image description. *International Journal Of Computer Vision*, 45(2):83–105.

Keselman, Y. and Dickinson, S. (2005). Generic model abstraction from examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1141–1156.

Kimia, B., Tannenbaum, A., and Zucker, S. (1995). Shapes, shocks, and deformations i: The components of two-dimensional shape and the reaction-diffusion space. *International Journal Of Computer Vision*, 15(3):189–224.

Leibe, B. and Schiele, B. (2003). Analyzing appearance and contour based methods for object categorization. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, USA, June 2003*.

Li, F., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 28(4):594–611.

Lowe, D. G. (1985). *Perceptual organization and visual recognition.* Kluwer Academic Publishers, Boston.

Malik, J., Belongie, S., Leung, T., and Shi, J. (2001). Contour and texture analysis for image segmentation. *International Journal Of Computer Vision*, 43(1):7–27.

Marr, D. (1982). *Vision*. W. H. Freeman, New York.

Martin, D., Fowlkes, C., and Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549.

Mel, B. W. (1997). Seemore: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput*, 9(4):777–804.

Minsky, M. (1975). A framework for representing knowledge. In Winston, P., editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, New York.

Mojsilovic, A., Gomes, J., and Rogowitz, B. (2004). Semantic-friendly indexing and quering of images based on the extraction of the objective semantic cues. *International Journal Of Computer Vision*, 56(1-2):79–107.

Mokhtarian, F. and Mackworth, A. (1986). Scale-based description and recognition of planar curves and two-dimensional shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:34–43.

Mokhtarian, F. and Mackworth, A. (1992). A theory of multiscale, curvature-based shape representation for planar curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:789–805.

Mori, G., Belongie, S., and Malik, J. (2005). Efficient shape matching using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1832–1837.

Motoyoshi, I., Nishida, S., Sharan, L., and Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, 447(7141):206–209.

Nelson, R. and Selinger, A. (1998). A cubist approach to object recognition. In *Sixth International Conference on Computer Vision*.

Niethammer, M., Betelu, S., Sapiro, G., Tannenbaum, A., and Giblin, P. (2004). Area-based medial axis of planar curves. *International Journal Of Computer Vision*, 60(3):203–224.

Noton, D. & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, 171:308–311.

Novak, D., Batko, M., and Zezula, P. (2008). Web-scale system for image similarity search: When the dreams are coming true. *IEEE CBMI 2008, London*, pages 446–453.

Ogniewicz, R. and Kubler, O. (1995). Voronoi tessellation of points with integer coordinates: Time- efficient implementation and online edge-list generation. *Pattern Recognit.*, 28(12):1839–1844.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.*, 42(3):145–175.

Opelt, A., Pinz, A., and Zisserman, A. (2006). A boundary-fragment-model for object detection. In *European Conference on Computer Vision (ECCV 2006), Part II, LNCS 3952*.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, Massachusetts.

Palmer, S. E., Rosch, E., and Chase, P. (1981). Canonical perspective and the perception of objects. In Long, J. and Baddeley, A., editors, *Attention and performance IX*, pages 135–151. Erlbaum, Hillsdale, NJ.

Parent, P. & Zucker, S. W. (1989). Trace inference, curvature consistency, and curve detection. *IEEE Trabsactions on pattern analysis and machine intelligence*, 11:823–839.

Pelillo, M., Siddiqi, K., and Zucker, S. (1999). Matching hierarchical structures using association graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:1105–1120.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *J Exp Psychol [Hum Learn]*, 2(5):509–22.

Privitera, C. and Stark, L. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye-fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:970–981.

Rajashekar, U., Van der Linde, I., Bovik, A., and Cormack, L. (2008). Gaffe: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, 17:564–573. To Appear in: Transaction on Image Processing., 2008, URL: http://live.ece.utexas.edu/research/gaffe.

Rasche, C. (2005). *The Making of a Neuromorphic Visual System*. Springer, Berlin, Heidelberg, New York.

Rasche, C. (2007). Neuromorphic excitable maps for visual processing. *IEEE Transactions on Neural Networks*, 18(2):520–529.

Ravishankar Rao, A. & Lohse, G. L. (1993). Identifying high level features of texture perception. *CVGIP: Graphical Models and Image Processing*, 55:218–233.

Renninger, L. and Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, 44(19):2301–2311.

Rolls, E. and Deco, G. (2002). *Computational neuroscience of vision*. Oxford University Press, New York.

Rosch, E., Mervis, C., Gray, W., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Rosenfeld, A. and Pfaltz, J. (1968). Distance functions on digital pictures. *Pattern Recognition*, 1(1):33–61.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 29(3):411–426.

Shotton, J., Blake, A., and Cipolla, R. (2005). Contour-based learning for object detection, pp. 503-510. In *The 10th IEEE International Conference on Computer Vision (ICCV05)*.

Shotton, J., Blake, A., and Cipolla, R. (2008). Multi-scale categorical object recognition using contour fragments. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 30(7):1270–1281.

Siddiqi, K., Kimia, B., Tannenbaum, A., and Zucker, S. (1999). Shapes, shocks and wiggles. *Image And Vision Computing*, 17(5-6):365–373.

Standing, L., Conezio, J., and Haber, R. N. (1970). Perception and memory for pictures - single-trial learning of 2500 visual stimuli. *Psychonomic Science*, 19(2):73–74.

Sudderth, E., Torralba, A., Freeman, W., and Willsky, A. (2008). Describing visual scenes using transformed objects and parts. *International Journal Of Computer Vision*, 77:291–330.

Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man, and Cybernetics*, 8:460–473.

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381:520–522.

Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scene: The role of global features on object search. *Psychological Review*, 113:766–786.

Treisman, A. and Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychol Rev*, 95(1):15–48.

Tu, Z. and Zhu, S. (2006). Parsing images into regions, curves, and curve groups. *International Journal Of Computer Vision*, 69(2):223–249.

VanRullen, R. and Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research*, 42(23):2593–2615.

Vogel, J. and Schiele, B. (2007). Semantic modeling of natural scenes for content-based image retrieval. *International Journal Of Computer Vision*, 72(2):133–157.

Wang, J., Li, J., and Gio, W. (2001). Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963.

Wang, S., Stahl, J., Bailey, A., and Dropps, M. (2007). Global detection of salient convex boundaries. *International Journal Of Computer Vision*, 71:337–359.

Witkin, A. and Tenenbaum, J. (1983). On the role of structure in vision. In Beck, J., Hope, B., and Rosenfeld, A., editors, *Human and machine vision*, pages 481–543. New York: Academic Press.

Yuille, A., Fang, F., Schrater, P., and Kersten, D. (2004). Human and ideal observers for detecting image curves. *Advances in Neural Information Processing Systems*, 16:1459–1466.

Zhang, D. and Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37:1–19.

Zhong, B. and Liao, W. (2007). Direct curvature scale space: Theory and corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:508–512.

Zhu, S. and Yuille, A. (1996). Forms: a flexible object recognition and modelling system. *International Journal Of Computer Vision*, 20:187–212.

Zhu, S.-C. (1999). Stochastic jump-diffusion process for computing medial axes in markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:1158–1169.

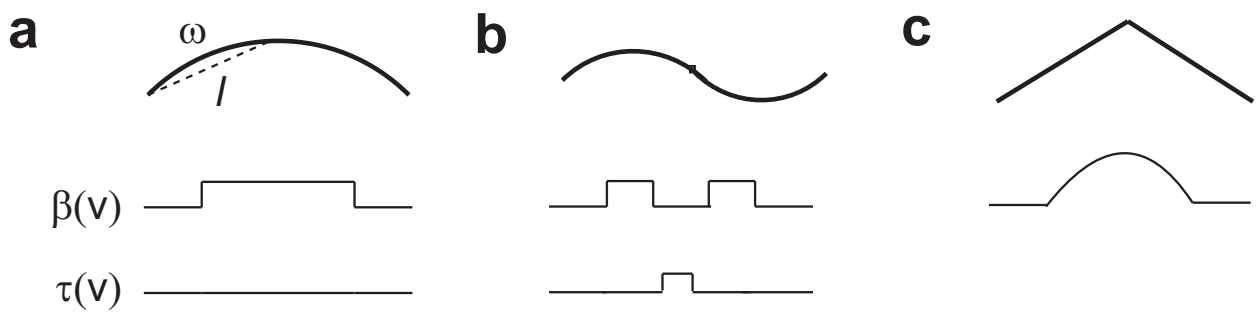

Figure 1: Signatures of elementary contour geometries. **a**. A chord (stick) of fixed length ω is iterated through the contour and the distance between the straight line ℓ and the segment determined. If the segment lies primarily on one side of the straight line, its maximal distance is attributed to the 'bowness' signature $\beta(v)$, if it lies on both sides the maximal distance is attributed to the inflexion signature $\tau(v)$ [v: arc length variable]. For an arc the bowness signature is a rectangular function (for a given window size). **b**. For an inflexion, the bowness signature consists of two short rectangular functions and a short rectangular function for the inflexion signature (also called signature blocks). **c**. Bowness signature for a L feature.

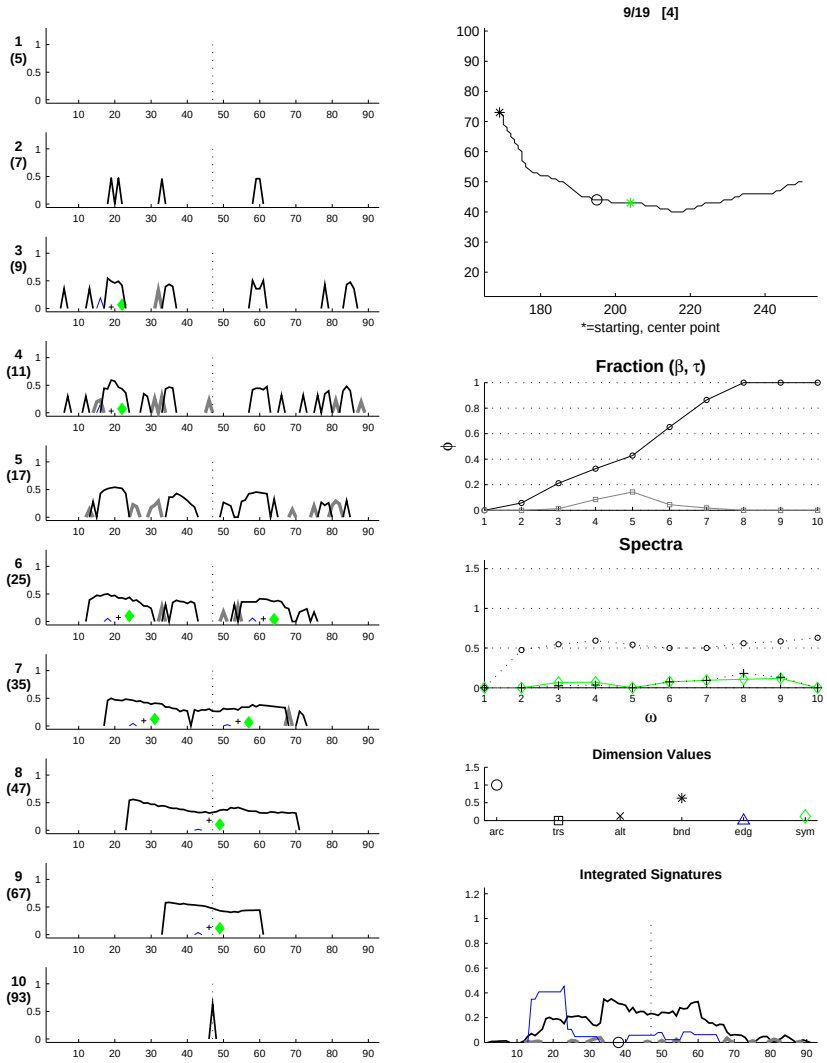

Figure 2: Local/global (LG) space of a wiggly arc. **Top right**: sample contour with starting and center points marked as asterisk. **Left column**: LG space: signatures $\beta(v)$ (black) and $\tau(v)$ (grey) for 10 different window sizes [x-axis= arc length variable v]. Signature block characteristics (determined for large ones only): blue marker= $\epsilon^{\sqcap}$; green diamond=$v^{\sqcap}$; plus sign=$\zeta^{\sqcap}$. **Fraction**: fraction ϕ of bowness and inflexion blocks per window size. **Spectra**: Green diamond: maximum of symmetry value; black circle: maximum β amplitude; plus sign: maximum of ζ. **Dimension Values**: arc, transition, alternation, bnd=curvature, edginess, symmetry. **Integrated Signatures**: bowness (black), inflexion (gray) and edginess (blue).

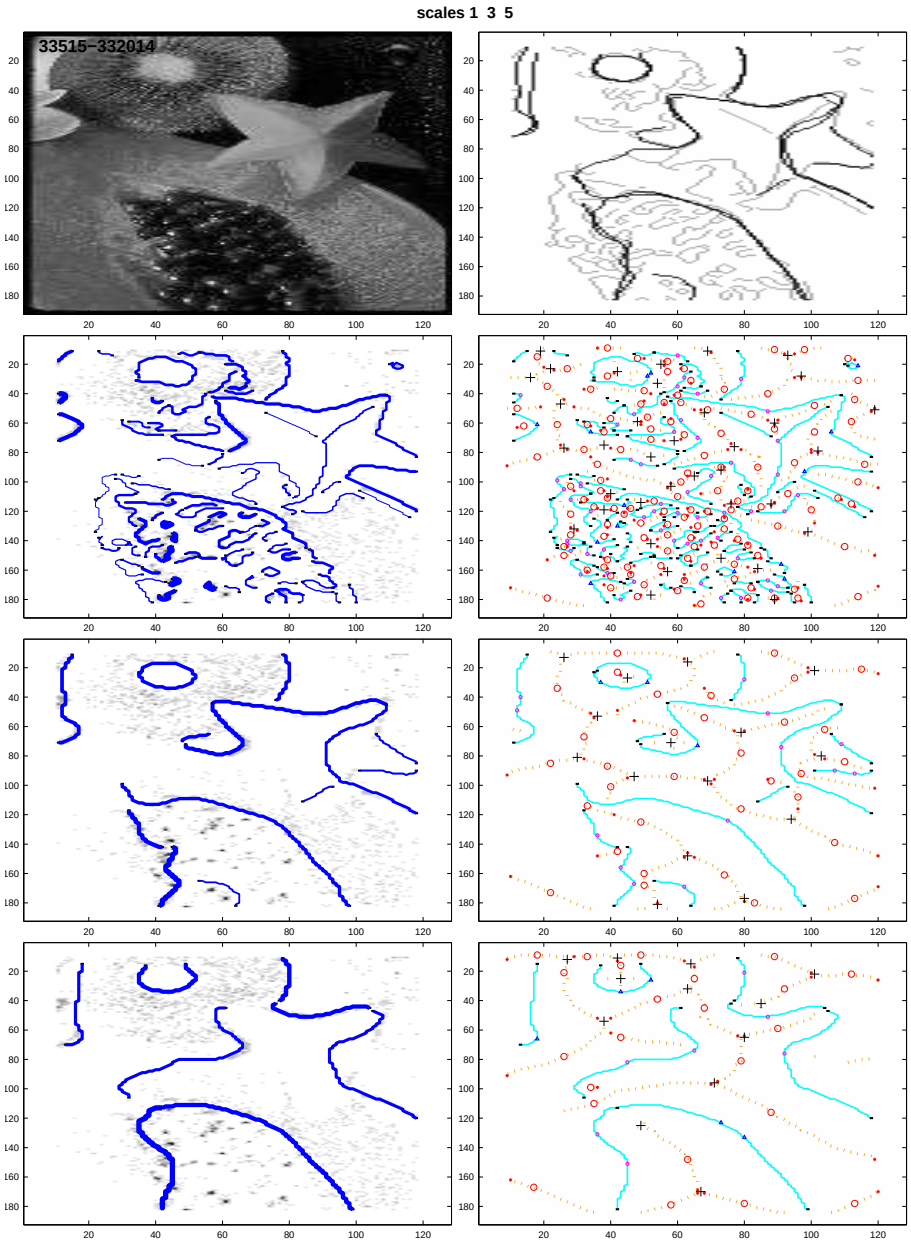

Figure 3: Summary of the decomposition ($\sigma = 1, 3, 5$). **Top right**. Contours of all three scales overlayed. **Left column**: Contours in blue, with line thickness corresponding to mean contrast (c_m). Gray-scale pixels represent output of fuzziness (blob) filter. **Right column**: Contours in cyan with start- and end-point marked as black dots; highest curvature marked by blue triangles or magenta circles; sym-axes orange dotted with red circle marking p_{fx} and red dot marking s_2. +: intersections of sym-axes.

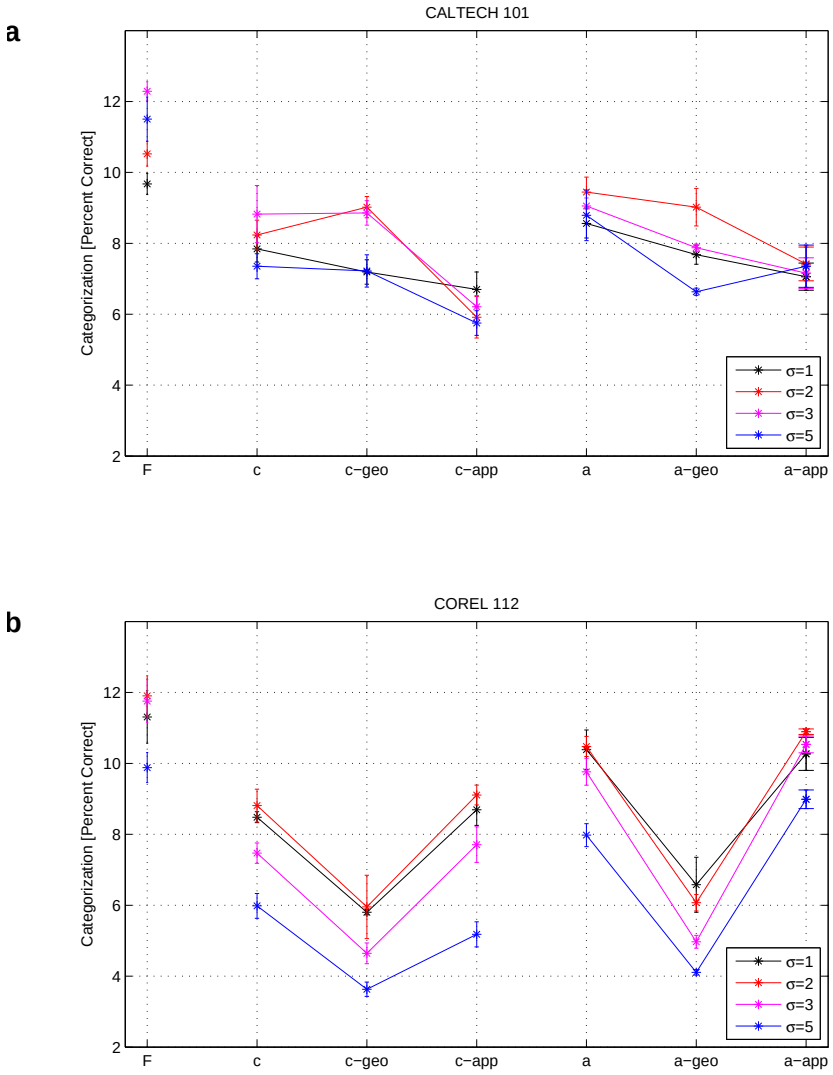

Figure 4: Categorization results of histogram matching for partial dimensionality for the Caltech (**a**) and Corel (**b**) set. 'F' full dimensionality (280 dimensions); 'c' and 'a': complete contour and area descriptors (130 and 150 dimensions); 'c-geo' and 'a-geo': geometrical dimensions only (90 dimensions each); 'c-app' and 'a-app': appearance dimensions only (40 and 60 dimensions) for 4 different spatial scales (σ=1, 2, 3, 5). Error bars denote standard error of 3-fold cross validation.

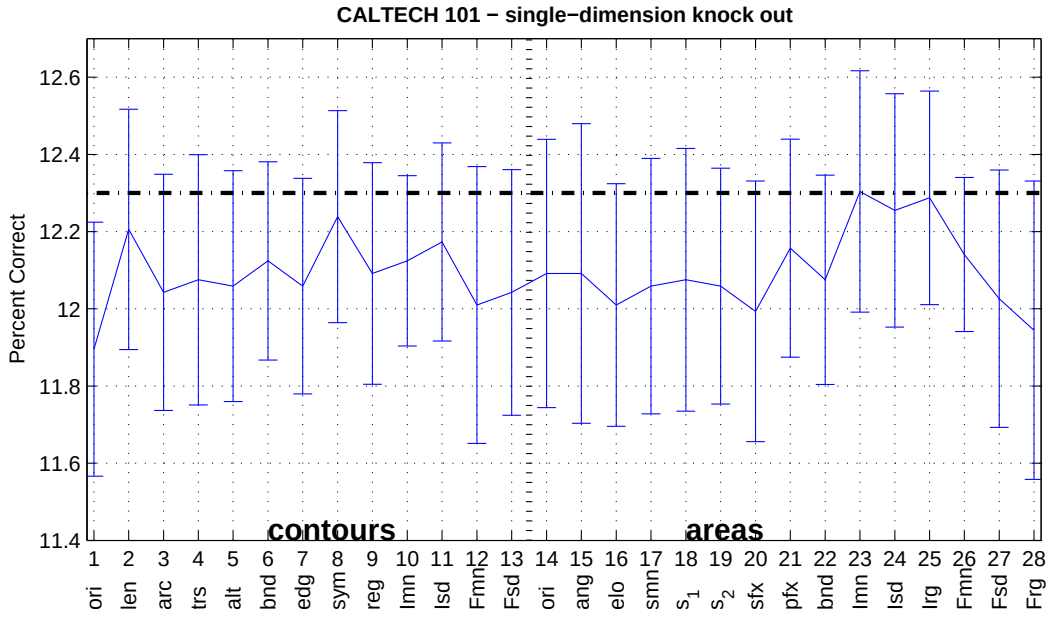

Figure 5: Categorization performance for single-dimension knock out for the Caltech collection (270 dimensions each). The performance for full dimensionality (280 dimensions) is indicated as black dashed line at 12.3 percent. Error bars denote standard error of 3-fold cross validation.

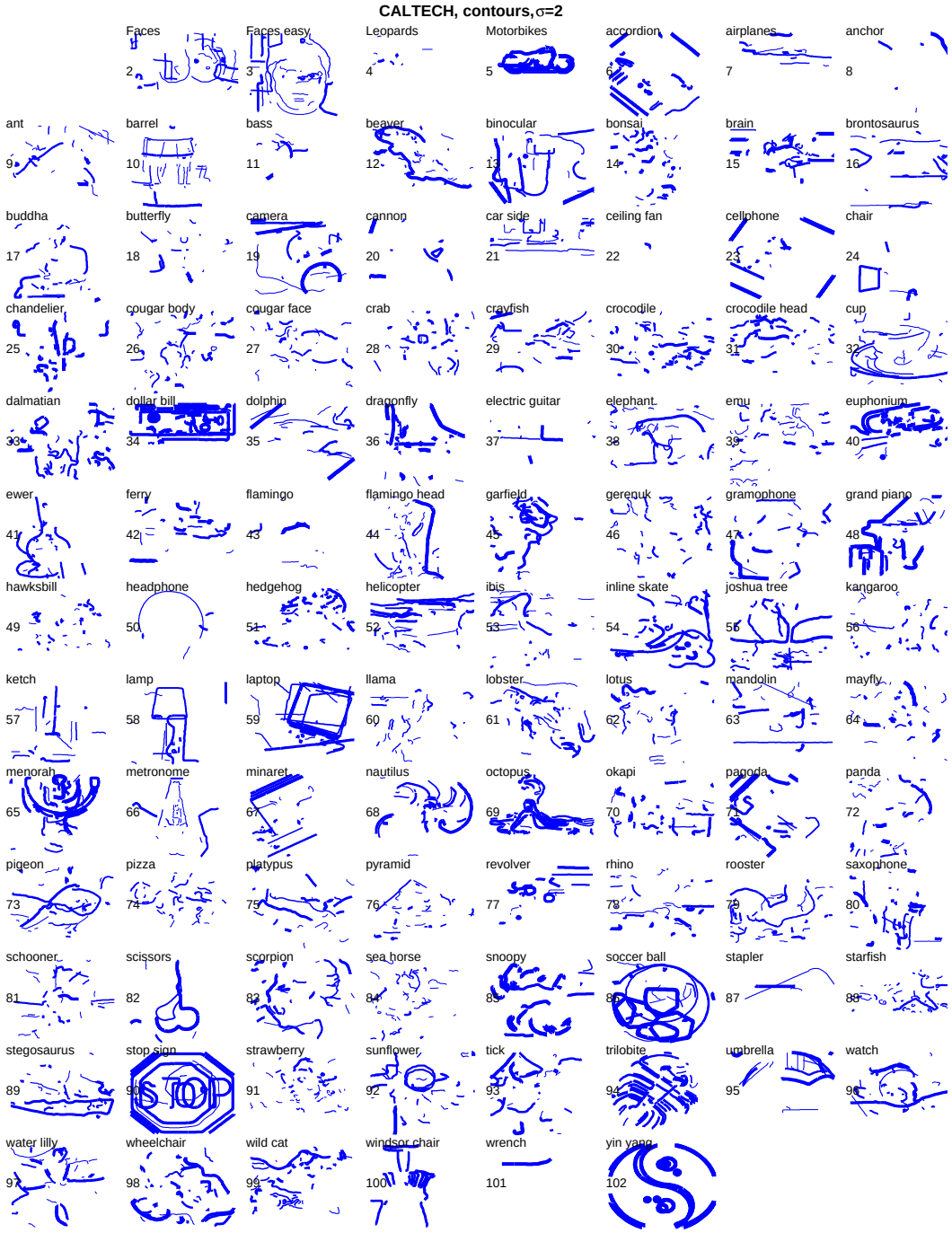

Figure 6: Category-specific contours from spatial scale $\sigma = 2$ for all 101 categories of the Caltech collection. Contours were obtained by image sorting and selecting those, which preferably found their own category images (average 3-8 percent, for those showing any specificity at all; maximum percentage larger than 50 percent, see also figure 8). The contour thickness corresponds to the average contrast value c_m.

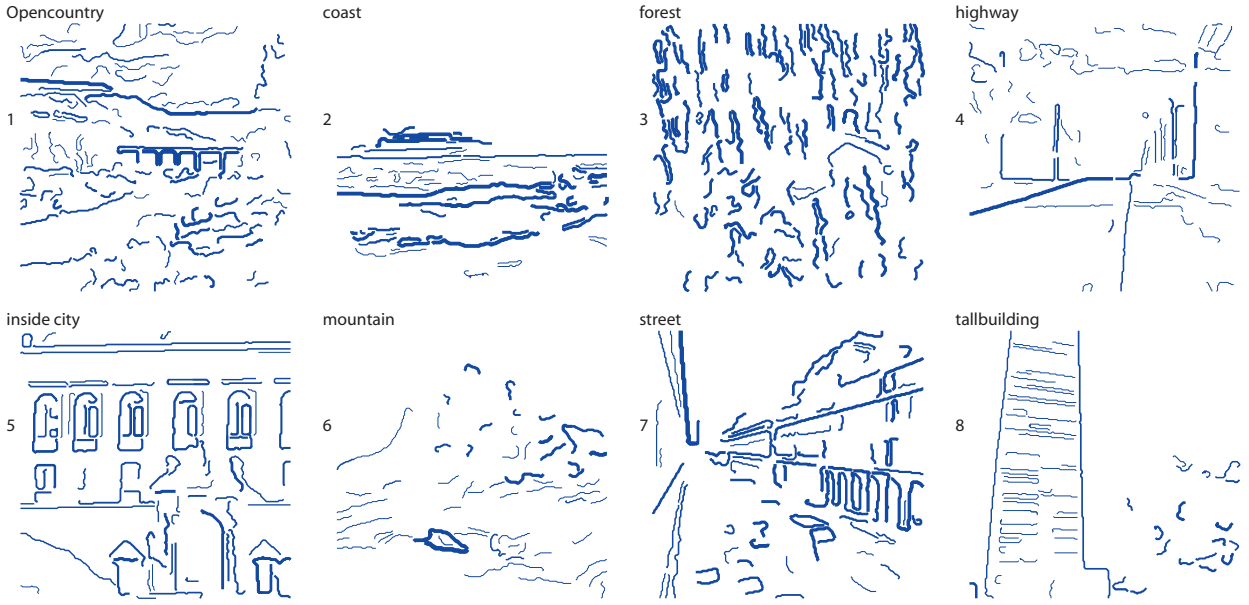

Figure 7: Category-specific contours from spatial scale $\sigma = 2$ for the Urban&Natural collection. See previous figure for more details.

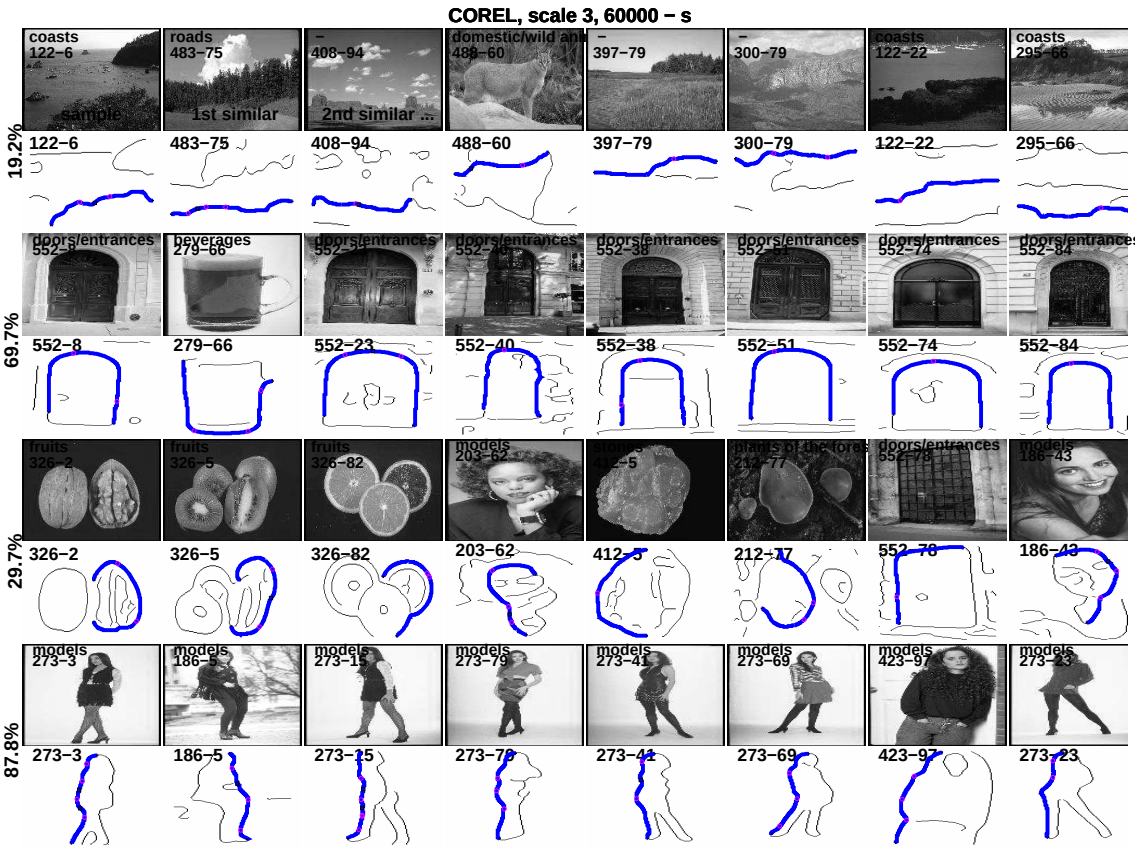

Figure 8: Similarity-based contour search for all contours of the entire Corel collection (60000 images) for $\sigma = 3$. The contour of the first image in each row is the selected sample contour, the remaining images in each row contain the most similar contours. The percentage on the left denotes correct basic-level categorization for the first 99 similar images.