RAID: A Relation-Augmented Image Descriptor

Paul Guerrero
KAUST, University College London

Niloy J. Mitra
University College London

Peter Wonka
KAUST

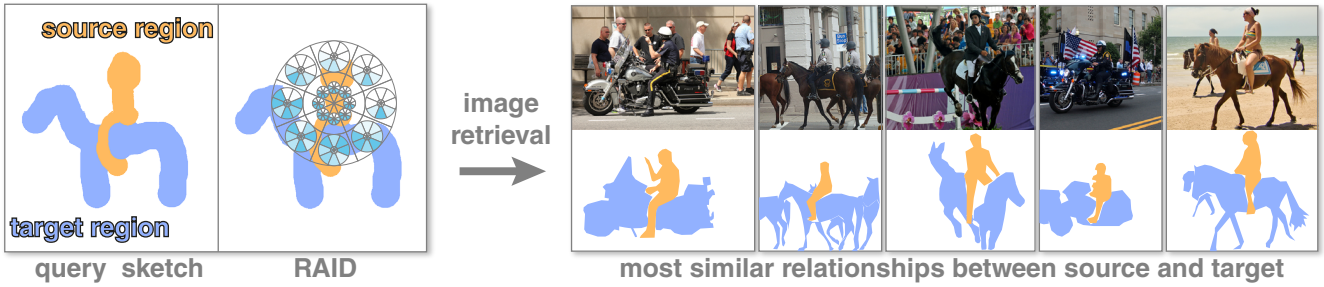


Figure 1: *We propose a novel descriptor called* RAID *to describe the spatial relationship between image regions. This descriptor enables retrieval with queries based on complex relationships between regions, such as the 'riding' relationship between the orange source and the blue target region. In this example, the user sketched two regions (left) and* RAID *retrieved images as shown (right).*

Abstract

As humans, we regularly interpret scenes based on how objects are *related*, rather than based on the objects themselves. For example, we see a person *riding* an object X or a plank *bridging* two objects. Current methods provide limited support to search for content based on such relations. We present RAID, a relation-augmented image descriptor that supports queries based on inter-region relations. The key idea of our descriptor is to encode region-to-region relations as the spatial distribution of point-to-region relationships between two image regions. RAID allows sketch-based retrieval and requires minimal training data, thus making it suited even for querying uncommon relations. We evaluate the proposed descriptor by querying into large image databases and successfully extract non-trivial images demonstrating complex inter-region relations, which are easily missed or erroneously classified by existing methods. We assess the robustness of RAID on multiple datasets even when the region segmentation is computed automatically or very noisy.

Keywords: spatial relationships, image descriptors, relation-based query, sketch-based query, image retrieval

Concepts: •**Computing methodologies** → **Shape analysis;**

1 Introduction

Detecting, encoding, and synthesizing relationships between objects is critical for many shape analysis and scene synthesis tasks. Handling even simple relations like 'on top of,' 'is next to,' or 'is touching' has been shown to be very useful for scene understanding [Liu et al. 2014], structuring raw RGBD images [Shao et al. 2014], realistic scene synthesis [Fisher et al. 2012; Chen et al. 2014], object retrieval [Fisher et al. 2011], etc. Recently, more advanced relationship descriptors like IBS [Zhao et al. 2014] and ICON [Hu et al. 2015] have demonstrated the value of capturing intricate relations via proxy objects [Zheng et al. 2014], or with human agents [Kim et al. 2014; Fisher et al. 2015].

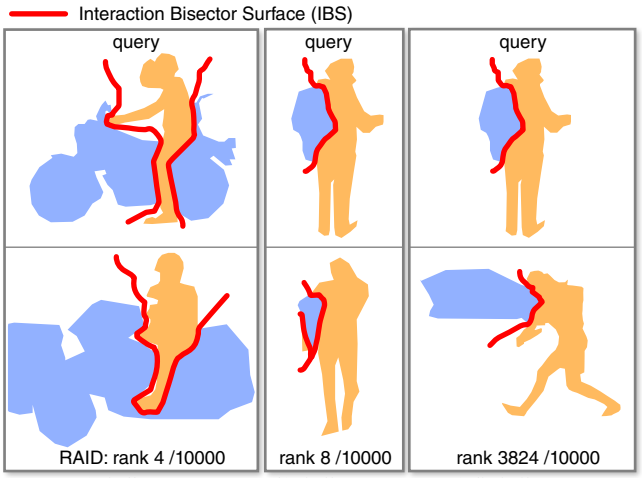
We present RAID, a relation-augmented image descriptor, to encode relationships between two regions in space. To be successful, a relationship descriptor should meet several expectations. Most importantly, a successful relationship descriptor should be discriminative enough to distinguish between simple relationships like 'on top of,' but also between complex relationships not treated by previous work, e.g., 'is enclosed by,' or 'is leaning on.' Additionally, the descriptor should be compact, enable fast comparisons between two relationships, and be robust to noise in the segmentation.

There are two strategies to obtain such a descriptor. First, a descriptor can be automatically learned, e.g., using convolutional neural networks. However, this requires a large amount of training data, possibly several thousands of labeled images. While millions of labeled images exist, some with object segmentations or object class labels, rarely do they come with relationship labels. Since obtaining such large quantities of data from scratch seemed impractical we opted against this approach. Second, a descriptor can be designed. Based on our experiments, Shape Contexts [Belongie et al. 2002] are the best available relationship descriptors for two-dimensional regions that describe simple *point-to-region* relationships such as 'below' or 'adjacent.' However, in a complex relationship between two regions, these simple point-to-region relationships usually vary over a region. For example, in Figure 3, the head of the man is above the bench, while his feet are below. The key idea behind RAID is to capture the spatial *distribution* of such simple point-to-region relationships to describe more complex relationships between two image regions. Capturing the relationship distribution over the whole region, not just over separating surfaces like the IBS descriptor, makes our descriptor robust to topological noise and increases its discriminative power. Figure 2 presents a comparison.

RAID enables relationship-based retrievals, which is fundamentally different from retrievals based on keywords, color histograms [Pentland et al. 1996; Arnold et al. 2000], object sketches [Eitz et al. 2009a; Cao et al. 2011], or using a rough composition guidance [Hu et al. 2013]. Specifically, given a single desired composition of abstract regions, with or without labels, as an exemplar of a target relationship, RAID retrieves images exhibiting this relationship. As regions, RAID uses either automatically segmented regions [Zheng et al. 2015], or available segmentation and labelling information (cf., [Malisiewicz and A. 2009]).

SIGGRAPH '16 Technical Paper, July 24-28, 2016, Anaheim, CA,
ISBN: 978-1-4503-4279-7/16/07
DOI: http://dx.doi.org/10.1145/2897824.2925939

ACM Reference Format
Guerrero, P., Mitra, N., Wonka, P. 2016. RAID: A Relation-Augmented Image Descriptor. ACM Trans. Graph. 35, 4, Article 46 (July 2016), 12 pages. DOI = 10.1145/2897824.2925939 http://doi.acm.org/10.1145/2897824.2925939.



(a) similar RAID, but different IBS (b) similar RAID, but different IBS (c) dissimilar RAID, but similar IBS

Figure 2: *Comparison of RAID and Interaction Bisector Surfaces (IBS). We show three example queries where IBS is indicated with red curves. Since our descriptor uses information from the entire source region (orange), our results are different from IBS. For queries (a) and (b), we show results that are similar in RAID, but different in IBS, mainly due to topological differences. Query (c) shows a result where the separating surfaces are similar, but the remainder of the regions is different. Note that (a) and (b) are false negative IBS results, whereas (c) is a false positive result for IBS.*

Such a tool immediately allows an artist to search for a particular scene configuration for inspiration, or a media creator to seek images with a particular assembly of objects. For example, the artist may ask for content with 'man *riding* a horse,' or 'man *standing next* to a horse,' or more generally 'man *riding* any object.' In a way, RAID enables querying by *verbs* relating image segment names by associating a particular descriptor with each such verb. The retrieved images can then be used to guide edit propagation [Berthouzoz et al. 2011; Yücer et al. 2012] by constraining edits to have a given relationship to the edited region, improve library-driven image synthesis [Hu et al. 2013] by returning more relevant regions from the library, or enhance image completion [Hays and Efros 2007; Huang et al. 2013] in context-dependent image regions.

We evaluate query performance as well as classification rates for RAID and compare against Shape Contexts as a baseline. We quantitatively measure performance as the precision of query results in a large dataset consisting of 10000+ images. Classification performance is tested on two smaller datasets, a synthetic dataset containing 164 images and a set of 75 images collected from the web. Additionally, we evaluate the robustness of our method to noisy inputs and compare the performance on manually segmented images versus automatically segmented images via two user studies. Results show that our method is able to successfully describe complex relationships with a clear improvement over Shape Contexts, is robust to boundary noise with average displacements up to 20% of the largest image dimension, and also performs well on automatically segmented regions, with a better performance than Shape Contexts applied to *manually* segmented regions.

To summarize, our main contribution is a method to encode complex relationships between 2D image regions in a simple descriptor that can be used to query large databases efficiently and does not need large volumes of training data. Further, RAID provides valuable insights for extensions to directly capture relationships in 3D.

2 Related Work

Most research on spatial relationships between image regions has been done in the field of content-based image retrieval. These methods usually focus on describing the composition of *all* regions in an image and use relatively simple models for pair-wise relationships. The survey of Bloch [2005] gives a good overview of early methods that include statistics over distances or directions (although not both) between points in both regions. These methods do not attempt to describe complex relationships or capture a spatial distribution of relationships. More recent approaches can be classified by the type of models they employ, as outlined in the following paragraphs.

Shape descriptors. Several shape descriptors have been proposed over the last two decades. Surveys can be found in [Zhang and Lu 2004; Kazmi et al. 2013]. Some region-based shape descriptors can be adapted to describe the simple relationship between a point and an image region. These include polar and square shape matrices [Goshtasby 1985; Flusser 1992], moment-based shape descriptors [Teague 1980; Celebi and Aslandogan 2005] and Shape Contexts [Belongie et al. 2002]. In this work, we describe a novel descriptor for complex relationship between two image regions. We use Shape Contexts [Belongie et al. 2002] as a baseline shape descriptor in our performance evaluations. Recently, two descriptors for geometric interactions of 3D objects have been proposed, IBS [Zhao et al. 2014] and ICON [Hu et al. 2015]. IBS characterizes the negative space between the objects as a subset of the Voronoi diagram defined between the objects. This has been used successfully to characterize the interaction between 3D objects. However, there are three main differences between IBS and our descriptor. First, the subset of the Voronoi diagram used in IBS can be sensitive to topology changes and topological noise. Image regions are prone to topological noise and similar interactions between image regions often exhibit different topology (see Figure 2), making such a descriptor less suitable. Second, only the separating surfaces (i.e., medial surfaces) are captured; the remainder of the interacting regions is not represented. Third, the histogram binning used for IBS features does not preserve information about the spatial distribution of these features. In contrast, RAID is robust to topological variations and captures the spatial distribution of features over the entire shape of image regions, increasing its discriminative power. We provide a quantiative comparison to IBS in Section 6. ICON, which generalizes IBS, suffers from the same limitations.

Sketch-based retrieval. One application of our method shown in Figure 1 is sketch-based image retrieval. Several methods have been proposed [Eitz et al. 2009b; Eitz et al. 2011; Eitz et al. 2012] that either retrieve 3D objects or images based on sketches. However these methods search for properties of a single object shown in the sketch. In our system, *two* regions are sketched and we search for regions with similar relationships. In Sketch2Photo [Chen et al. 2009] and Sketch2Scene [Xu et al. 2013] compositions of objects are sketched, but objects are retrieved individually, or based only on simple relationships that describe the context of the sketched object, e.g., the relative bounding box position and vertical supports.

Scene understanding and machine learning. An important part of scene understanding is to accurately identify the relationship between scene objects. Several methods tackle this challenge by creating models of region relationships. Malisiewicz and Efros [2009] encode the spatial context of image regions in a graph. Features used in the spatial context are the amount of overlap, relative displacement, relative scale and relative height between two regions. Kulkarni et al. [2013] use one specialized detector for each of their 16 simple relationship classes, such as 'above', 'on', and 'near'. Adding an additional class requires implementing an additional detector. Our approach describes more complex relationships and provides a single data-driven descriptor for all relationship classes.

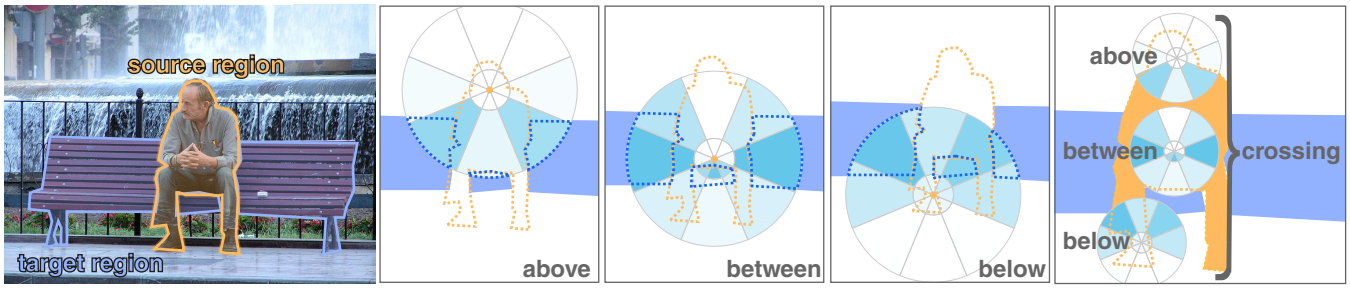


Figure 3: *Simple and complex relationships between the man and the bench shown on the left. We can identify several simple relationships between points in the source region (man) and the target region (bench). The relationships of each point are described by a polar histogram, with each bin colored according to the percentage of overlap with the target region. Some points are above the bench, some are below, and some are in between the bench segments. When looking at the spatial distribution of these simple relationships, we can infer the more complex 'crossing' relationship between source and target region.*

Data-driven methods have recently gained popularity in research on scene understanding, mainly using deep neural networks trained either directly on images [Karpathy and Li 2015] or on features extracted from images [Jansen et al. 2015]. These methods achieve impressive results, but they require extremely large training datasets and would not produce reasonable results with a single example relationship. Additionally, given a set of labeled training data, only a fixed set of categories can be learned and training would need to be repeated with a new set of annotations for additional categories. Manually annotating relationships in image databases is more difficult than annotating objects, due to the quadratic number of relationships in an image, the fact that many object pairs might need multiple labels, and the difficulty of finding a useful and unambiguous set of categories for relationships. Some recent databases contain region relationships [Chao et al. 2015; Krishna et al. 2016] that could be used as training data. However, significant effort is involved in creating these databases and due to the number of examples needed for each relationship category, only a fixed number of categories can be learned from these datasets. Krishna et al. [2016] also reported some difficulties in generating an unambiguous set of relationship categories. Additionally, these databases capture *semantic* relationships, which are less suitable for learning the geometric relationships captured by RAID. Our method does not need to be trained on large datasets, enabling example-based queries for arbitrary relationships. We believe that our contribution is orthogonal to learning in general and that RAID could be useful for supervised and unsupervised learning, either as a method to augment training data by automatically annotating region relationships, or directly as a feature vector.

String-based relationship models. A different line of research uses strings to describe the spatial layout of regions in an image [Wang 2003; Hsieh and Hsu 2008]. These methods project the image regions to the x- and y-axes of the image and record the starting point and end point of each projected region in two strings: one for the x-axis and one for the y-axis. This provides a compact representation of the region layout. However, a lot of information is lost during the projection to the image axes, resulting in a less discriminative description of relationships (for example, 'surrounded' cannot be distinguished from 'in a concave').

Point-based relationship models. One class of methods represent each image region as a single point, usually the centroid or bounding box center. As a consequence, only simple relationships, such as the distance [Ko and Byun 2002] or the direction [Lee and Hwang 2002; Lan et al. 2012; Huang et al. 2014] between the representative points, are captured (including relationships like 'below' and 'above'). A richer description of region relationships is presented by Zhou et al. [2001], based on the directional interval subtended by one region relative to the centroid of the other region. Complex relationships between two regions, however, can not be captured since one of the regions is still represented as a point.

Adjacency-based relationship models. Several methods [Chandran and Kiran 2003; Badadapure 2013] describe the layout of image regions as a graph, where nodes correspond to regions and edges connect adjacent regions. Region layouts can be compared efficiently using techniques from graph theory. Again, no attempt is made to describe complex relationships or the spatial distribution of relationships over a region. Similar to our paper, Hu et al. [2013] try to find matching regions in a large image library based on inter-region relationships. Relationships between adjacent image regions are described by a histogram of the relative locations between border pixels in a small two-pixel neighborhood. This allows capturing simple relationships between adjacent regions like 'above' or 'below'. In contrast, our approach describes a spatial distribution of relationships, enabling us to capture more complex relationships between image regions that do not need to be adjacent.

Image search methods using spatial distribution of different attributes have also been proposed. Such attributes include different encoding of spatial color distribution [Chua et al. 1997; Smith and Chang 1996; Ooi et al. 1998; Wang and Hua 2011], spatial distribution of text labels [Xu et al. 2010] (called context map), visual composites [Sadeghi and Farhadi 2011; Lan et al. 2013], spatial relationships among people [Choi et al. 2009], or self-similarity in regions [Shechtman and Irani 2007] etc. However, these methods do not support complex relationships between regions in the image or the spatial distribution of such relationships.

3 Relationships Between Image Regions

Here, we provide a definition of spatial relationships between two image regions and give several examples of such relationships. While most images that we consider are two-dimensional projections of three-dimensional scenes, our goal is to describe the two-dimensional composition of image regions rather than inferring a three-dimensional layout of the scene and then analyzing relationships in three dimensions. The advantage of this design choice is that the approach is a lot more robust, because inferring three-dimensional layouts from a single image is a challenging and underdetermined problem.

We can identify several classes of relationships that are commonly encountered in images. Examples are 'between', 'bridging', 'arching', 'crossing', as shown in Figure 4. We can observe, that most of the relationships are asymmetrical. For example, if region A is to the left of region B, then region B is to the right of region A. It is therefore necessary to distinguish between the two regions involved in a relationship. We call the first region in a relationship the source region, and the second region the target region.

For the purpose of this paper, we use a simple categorization to distinguish between simple and complex relationships. A simple relationship is one that exists for source points as well as source

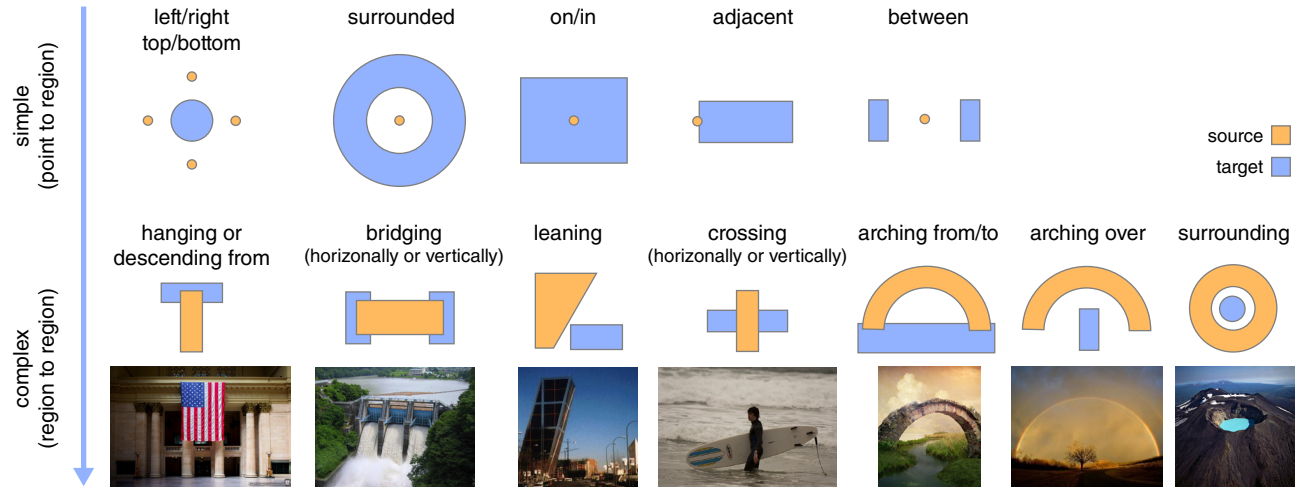


Figure 4: *Classes of spatial relationships between two-dimensional image regions. We distinguish simple relationships (top row) and complex relationships (middle row) between the orange source region and the blue target regions. Example images are shown below each complex relationship.*

regions. For example, both a point and a region can be surrounded by another region or can be above another region. A complex relationship can only exist for source regions larger than a single point. For example, only a region and not a point can surround another region or bridge two regions. Hence, the 'surrounded' and 'above' relationship are simple while the 'surrounding' and 'bridging' relationships are complex. In Figure 4, examples of simple relationships are shown on top and examples of complex relationships are shown in the middle row.

While there are several well-established methods to describe simple relationships, most importantly Shape Contexts [Belongie et al. 2002], in this paper we set out to design a descriptor to describe complex relationships as well as simple ones.

We use the following definitions:

The domain I of an image is a rectangular subset of $\mathbb{R}^2$. An image region A is defined as a subset of I. A labeling of an image region is a function $l : \mathbf{A} \rightarrow L$, where $\mathbf{A}$ is the set of all image regions and L is a label set.

A relationship class is a function that assigns a binary class membership to a pair of regions:

$$C_x(S,T) = \begin{cases} 1 & \text{if } S \text{ is in relationship } x \text{ with } T \\ 0 & \text{otherwise.} \end{cases}$$

Note that the same pair of regions can be members of multiple relationship classes. Further, in some datasets, labeled regions are disjunct (e.g., the COCO dataset) while some other data sets allow for overlaps between labeled regions (e.g. the synthetic and web datasets). In the next section, we propose a novel descriptor that is able to encode complex relationships.

4 The RAID Descriptor

The aim of our descriptor is to provide a numerical description of the relationship between a given source region S and a given target region T. We build on the fundamental observation that a complex relationship between S and T can be characterized by the relationship of each point in S to each point in T. Our approach to build the descriptor was therefore to first describe the relationship of each point in S to the region T separately. Afterwards, the problem becomes finding a suitable way to aggregate all individual point to region descriptors. In the following, we describe our solution to encode the distribution of point relationships over S.

A point relationship is described by a two-dimensional histogram $H(\mathbf{s})$ of the distance and direction between a source point $\mathbf{s}$ and each point $\mathbf{t}$ in the target region, similar to Shape Contexts [Belongie et al. 2002]:

$$H_{ij}(\mathbf{s}) = \frac{1}{a_{ij}} \int_{\Phi_i} \int_{R_j} \mathbf{1}_T(\mathbf{s} + (r\cos\phi, r\sin\phi)^{\mathrm{T}})\; r\; dr\; d\phi, \quad (1)$$

where Φ_i and R_j are respectively the angular and radial intervals of bin (i,j), and $\mathbf{1}$ is the indicator function. Each bin is normalized by the bin area a_{ij}. We call this histogram the *point histogram*. The center image in Figure 5, shows an example of two regions in the bridging relationship for points $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$. Basically, a histogram bin will contain a value H_{ij} corresponding to the fraction of its area covered by region T.

The distribution of point relationships over the source region is then encoded by a second histogram $\hat{\mathcal{H}}^S$ over the individual point histograms, resulting in a four-dimensional histogram:

$$\hat{\mathcal{H}}^S_{ijkl} = \frac{\int_{\Phi_k} \int_{R_l} (\mathbf{1}_S H_{ij})(\mathbf{c} + (r\cos\phi, r\sin\phi)^{\mathrm{T}})\; r\; dr\; d\phi}{\int_{\Phi_k} \int_{R_l} \mathbf{1}_S(\mathbf{c} + (r\cos\phi, r\sin\phi)^{\mathrm{T}})\; r\; dr\; d\phi}, \quad (2)$$

where $\mathbf{c}$ is the centroid of the source region. The rightmost image in Figure 5 shows an illustration of the 4D histogram. The denominator normalizes each bin by the intersection of the bin area with the source region. Bins with zero intersection (bins outside the source region) are assigned the value of the point histogram at the closest point of the source region. This gives a distribution of point histogram values over the source region that is not biased by bin areas or their coverage of the source region and effectively factors out the dependence of the histogram on the exact shape of the source region. The result is a spatial distribution of *relative* point positions between source and target regions. Finally, we perform a histogram normalization:

$$\mathcal{H}^S_{ijkl} = \frac{\hat{\mathcal{H}}^S_{ijkl}}{\sum_{ijkl} \hat{\mathcal{H}}^S_{ijkl}}. \quad (3)$$

We call this histogram the RAID descriptor. Similar to the SIFT descriptor [Lowe 2004], the RAID descriptor is a histogram of histograms, but RAID encodes directions and distances to a target region while SIFT encodes gradient orientations.

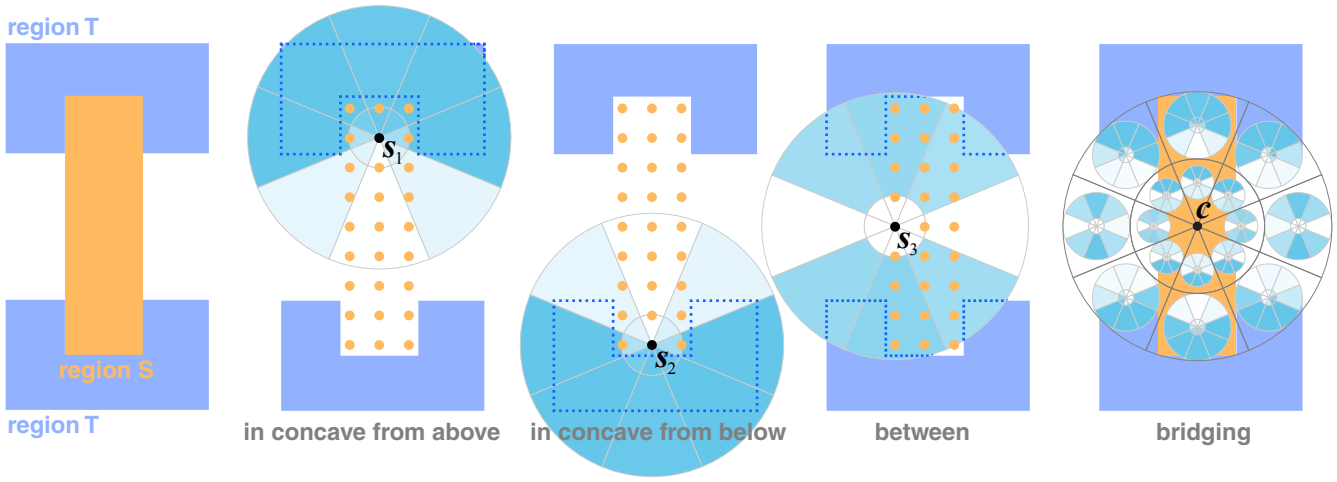


Figure 5: *The* RAID *descriptor of the relationship between two image regions S and T. In this example, region S 'bridges' region T vertically. Simple relationships between individual points $\mathbf{s}$ in S and region T are described by histograms of relative distance and direction from $\mathbf{s}$ to points in T: $\mathbf{s}_1$ and $\mathbf{s}_2$ are in a concave of T, while $\mathbf{s}_3$ is between T. More complex relationships between regions S and T are characterized by the distribution of simple relationships over S, which we capture in a histogram of simple relationships (rightmost image). In the 'bridging' relationship shown here, points like $\mathbf{s}_3$ that are between T are added to bins closer to the centroid $\mathbf{c}$, while points like $\mathbf{s}_1$ and $\mathbf{s}_2$ that are in a concave part of T contribute to bins further above and below. Note that the histograms in each bin on the right are scaled down for illustration only; they have the same size as the histograms shown in the center images.*

5 Implementation

In our implementation, we assume that image regions are given as polygons. The integral for the point histogram in Equation 1 can then be computed accurately and efficiently by constructing the Boolean intersection between the target region polygons and a set of polygons representing each bin of the point histograms. An efficient and robust implementation of this operation is available in the Boost polygon library [Boo 2015].

The integral in Equation 2 involves finding a point histogram for each source point. An analytical solution is not feasible, we therefore resort to an approximation. First, point histograms are computed at a regular grid of samples $\mathbf{s}$ inside the source region, by solving Equation 1 analytically as described above. As a good tradeoff between performance and accuracy, the density is chosen to be approximately $10000/a_I$, where a_I is the image area. Due to the limited sample density, directly accumulating these point histograms in the bins of the RAID descriptor would result in considerable aliasing, especially for smaller bins. Instead, we approximate the integral over a bin with a sum over all samples, weighted by a Gaussian kernel centered inside the bin:

$$\hat{\mathcal{H}}^S_{ijkl} = \frac{\sum_{\mathbf{s}\in\mathbf{S}} H_{ij}(\mathbf{s})\mathcal{G}(\mathbf{s}|\mathbf{c}+\mathbf{b}_{kl},\sigma^2)}{\sum_{\mathbf{s}\in\mathbf{S}} \mathcal{G}(\mathbf{s}|\mathbf{c}+\mathbf{b}_{kl},\sigma^2)}, \tag{4}$$

where $\mathbf{S}$ is the set of samples inside the source region, $\mathbf{b}_{kl}$ is the centroid of bin (k,l) relative to the histogram center and $\mathcal{G}(\mathbf{x}|\boldsymbol{\mu},\sigma^2)$ is an isotropic two-dimensional Gaussian with mean $\boldsymbol{\mu}$ and variance σ^2. The variance of the Gaussian is chosen so that the volume under the function equals the volume under the characteristic function of the bin. Note that this is a relatively coarse approximation, but it is efficient and works well as long as the shape of the bins is not too thin and elongated. As in Equation 3, the final discretized descriptor is then obtained through histogram normalization:

$$\mathcal{H}^S_{ijkl} = \frac{\hat{\mathcal{H}}^S_{ijkl}}{\sum_{ijkl}\hat{\mathcal{H}}^S_{ijkl}}. \tag{5}$$

In all our experiments, we set the maximum distance $r_{\max}$ for the outermost bin in the RAID descriptor to the maximum distance between the source region centroid and any other point in the source region. This ensures that the RAID descriptor covers the entire source region and effectively makes the descriptor scale-invariant. The maximum distance for the point histograms is set to the same value, meaning that an offset of $r_{\max}$ around the source region is captured by our descriptor. Our implementation uses 8 bins for both angular dimensions and 2 bins for both radial dimensions, giving a total of 256 bins. The descriptor geometry is shown in Figure 5. Images in the center show the size of bins (i,j) relative to the source region, the rightmost image shows the size of bins (k,l) (note that the histograms shown inside each bin (k,l) are scaled down for illustration only). Rotational invariance could be achieved by aligning the descriptor to the first principal component of the points in the source region. However, on many types of images, rotational invariance is not desirable (e.g. 'bridging horizontally' is different from 'bridging vertically'). We therefore keep the descriptor aligned to the x-axis of the image.

6 Evaluation and Applications

To evaluate the performance of our descriptor, we performed experiments on 10000 images of the Microsoft COCO dataset [Lin et al. 2014], a smaller synthetic dataset, and a small dataset of images collected from the web. The COCO subset contains a large variety of photographs that are suitable to evaluate the real-world performance of our method. However due to its large size, annotating every relationship to measure classification performance is not feasible. Instead, we perform image retrieval on this dataset and annotate the n best results of each query. This ground truth is used to evaluate the precision of our method. The synthetic dataset contains several abstract shapes and is small enough to exhaustively annotate all relationships. We evaluate precision as well as recall on this dataset. To measure classification performance on real images, we could take a small subsample of the COCO dataset. This would result in severe undersampling of the more uncommon relationship classes, however. Considering this, we collected a set of 69 images containing a balanced mix of relationship classes from the web. All datasets were finalized before starting our experiments.

To the best of our knowledge, there currently exists no descriptor that explicitly attempts to describe complex relationships between image regions. Most methods only describe simple relationships; that is, they describe relationships that can also be found between a point and a region. In the following evaluations, we compare our method to Shape Contexts [Belongie et al. 2002] and IBS [Zhao et al. 2014] adapted to 2D shapes. Shape Contexts are computed over the target region, are placed at the same center point as RAID and have a radius comparable to RAID. Since our descriptor uses histograms similar to Shape Contexts to describe simple relationships, this comparison also demonstrates how adding information about the distribution of simple relationships results in a description that is better suited for complex relationships.

Computational Complexity and Performance. Computing our descriptor has a complexity of $O(N_s N_b)$, where N_s is the number of sample points in the source region and N_b the number of bins in the point histogram. Since the number of bins is constant, the complexity is linear in the area of the source region. Our simple, single-threaded Matlab implementation requires approximately 0.13 seconds per descriptor on average. The COCO subset contains roughly 236000 relationships (24 relationships per image on average), which gives a total time of 8.5 hours for an exhaustive query on the entire dataset. However, specifying a label for the source or target region lowers the number of relationships by a factor of typically 4–5. Additionally, we can precompute the descriptors for the entire dataset, which requires about 510 MB of space. Querying the dataset then only requires computing the L_1 distances between the query descriptor feature vector and the feature vectors of the precomputed descriptors, which requires roughly 0.46 seconds in our Matlab implementation.

Image Retrieval. An interesting area of application for the RAID descriptor is image retrieval from large databases. Our method can extend the search capability of a system by enabling queries for given relationships, such as 'riding' or 'standing on'. In the following, we describe experiments we performed with different relationship queries on a dataset of 10000 images from the Microsoft COCO dataset [Lin et al. 2014]. In this dataset, image regions are annotated by labeled polygons. The set of labels is consistent throughout the dataset and the annotation quality is relatively high, which makes it a good choice for our experiments.

To specify a relationship query, we can either mark a pair of regions in an existing image, or create a pair of regions synthetically, for example by drawing two simple polygons. Given the pair of regions, we compare their RAID descriptor with the descriptors of the region pairs in all dataset images. We treat the descriptor values as feature vectors and compare them using the L_1 distance, which does not overly penalize single bins that have a high mismatch. In our experiments, we treat all target regions with the same label in an image as a single region. This also improves the robustness of the query, since the segmentation of an image into regions is often ambiguous (e.g., sometimes books in a shelf are annotated individually; sometimes a whole row of books is annotated as a single region) and regions might be subdivided by occluding objects. We can optionally filter a query by the label of the source or target region. For example, we can query for relationships where the source region has the label 'person'. The descriptor for a pair of query regions can also be stored and associated with a specific verb such as 'riding' or 'surrounding'. This allows future queries to be formulated as sentences consisting of a subject (the label of the source region), a verb (the stored descriptor) and an object (the label of the target region), such as 'chairs surrounding table' or 'person riding X', where X stands for any label. Since RAID is scale-invariant, results may contain relationships between small regions in the background. To filter out these less salient results, we remove source regions with an area below 1% of the image area from the result.

The ground truth for all retrieval results was created in several user studies. One study was conducted for each retrieval experiment described below. As a reference, we compare the results of each experiment with the result of RAID applied to noiseless, manually segmented regions. In each study, users were asked to compare the relationships of several region pairs returned by a query to the corresponding example region pair and rate them as 'different', 'similar', or 'somewhat similar'. Images of the two datasets being compared were displayed in randomized order and users did not have knowledge of which of the two methods generated each result. A total of 12 subjects participated in one or several of the studies and each result relationship was rated by at least three subjects. Since different methods have overlapping results, relationships often have more than three ratings. To get the ground truth score of a relationship in a given query, we average all ratings given by subjects.

Results of six queries are shown in Figures 6 and 7. The queries in Figure 6, as well as the first query in Figure 7 use images from the dataset as query regions. In these queries, we only search for source regions with the label 'person'. The remaining two queries use synthetic query regions and search for source and target regions of any label. In the bottom row of each figure, we provide the precision of the first n results of the query as a function of n. In the 'riding' query (Figure 6, first row), the source region contains an interesting distribution of simple relationships, including source points above and source points in between the target region. Our descriptor successfully finds regions with a similar distribution of simple relationships, while Shape Contexts and IBS also return many false positives that have a different distribution of simple relationships. Similar results can be observed on the 'carrying', 'standing on' and 'holding' relationships. Note how a similar distribution of simple relationships also corresponds to regions that are intuitively similar to the query. For the two synthetic queries, our method also returns more relevant results. In the 'surrounding' query, for example, our descriptor successfully reproduces the gap between source and target region, while Shape Contexts ignore the gap. IBS descriptors focus on the interacting borders of image regions, but they are sensitive to noise and slight variations of these borders and ignore the spatial distribution of feature values, resulting in a performance similar to Shape Contexts.

Classification Performance on the Synthetic Dataset. We performed additional evaluation on a small synthetic dataset containing 164 manually created images. Each image shows a single source- and a single target region. These region pairs were labeled manually with zero, one, or multiple labels from among the seven complex relationship classes shown in Figure 4 plus the 'surrounded' relationship. Of the 164 relationships, 97 are labeled with one or more relationship classes; the remaining relationships do not correspond to any of the classes. Since relationships can be part of multiple classes (e.g. a bridge may be arching between and bridging two shores), we use multi-label classification. More specifically, we split the multi-label classification into several independent binary classifications, one for each relationship class. Each binary classification is performed by a k-NN classifier based on the L_1 distance of the RAID descriptors. We set $k = 5$, so that the five closest relationships are used to determine the labels of a given relationship.

Results of a leave-one-out cross-validation of the classifier and a comparison to IBS and Shape Contexts are shown in Figure 8. Since Shape Contexts only capture simple relationships between a point and a region, they perform poorly with more complex relationships. Note, for example, the large number of relationships that were incorrectly classified as not corresponding to any class, shown in the last column of the confusion matrix. IBS performs better than Shape Contexts on synthetic data, but still misses relationships due to differences in the geometry of interacting surfaces. The lack of spatial information makes it difficult to discriminate

Figure 6: *Three relationship queries between the orange source regions and the blue target regions shown in the first column (green background). Source regions are set to be persons, while target regions may have any label. Results are shown for the* RAID *descriptor and Shape Contexts (SC). In each row, we show the* n *best results for the query shown in the first column. The bottom row shows the precision of the* n *best results of* RAID*, IBS and SC as a function of* n*. Note how the* RAID *descriptor finds regions that are intuitively more similar to the query relationship.*

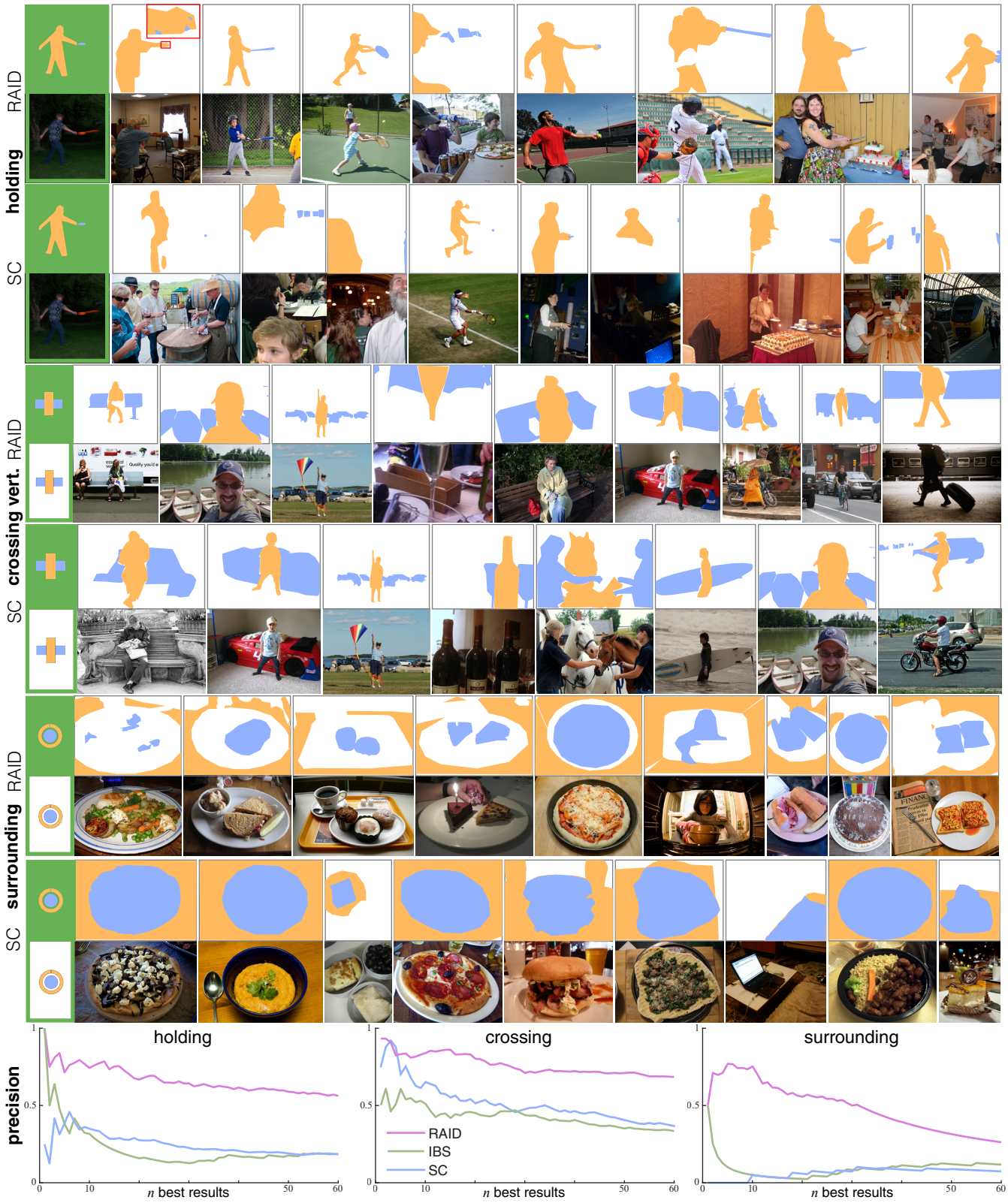


Figure 7: *Three additional relationship queries between the orange source regions and the blue target regions shown in the first column (green background). In the first query, source regions are set to be persons, while target regions may have any label. The second and third query were specified with a synthetic source and target region and relationships with any source and target labels were searched.*

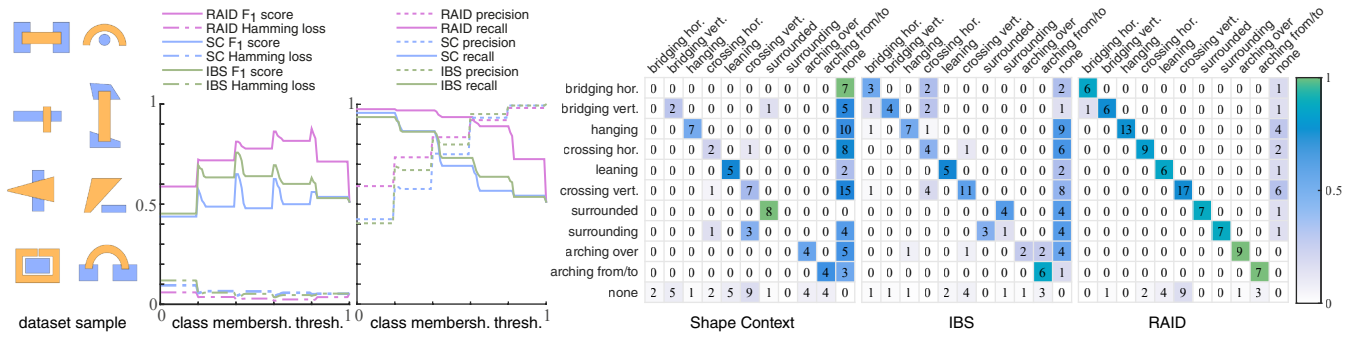


Figure 8: *Classification performance on the synthetic dataset and comparison to IBS and Shape Contexts. On the left, we show part of the dataset, followed by various performance measures at different class membership probability thresholds of the binary k-NN classifiers. Peaks in the F_1 scores and steps in the precision/recall scores are caused by sets of images with similar relationships being added correctly/incorrectly to the set of retrieved images as the threshold increases. On the right, confusion matrices are shown for each descriptor (rows correspond to actual classes, columns to predicted classes, colors are normalized by class size, while numbers show absolute values). Note that Shape Contexts are unable to detect some relationship classes, like 'bridging' or 'surrounding' and both SC and IBS miss many relationships.*

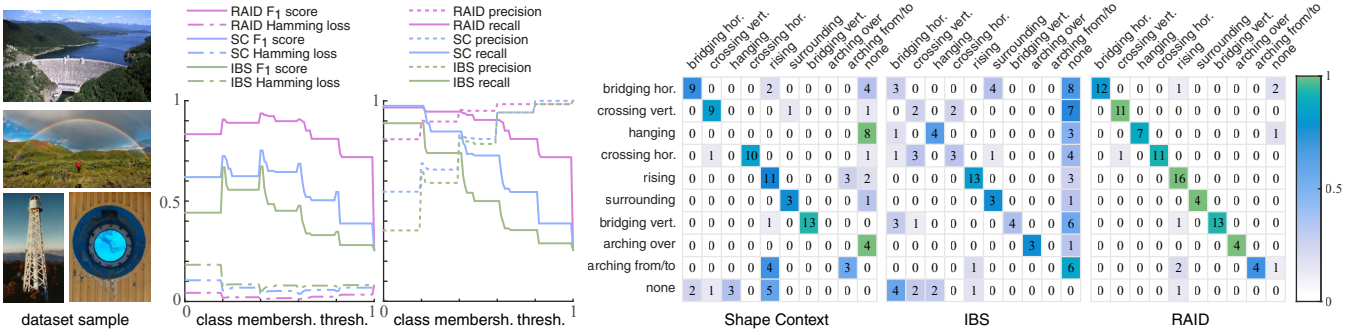


Figure 9: *Classification performance on the web image dataset and comparison to IBS and Shape Contexts. Four images of the dataset are shown on the left, each contains at least one of the relationship classes. In the center we show various performance measures at different class membership probability thresholds of the binary k-NN classifiers. On the right, the confusion matrices are shown for each descriptor (rows correspond to actual classes, columns to predicted classes, colors are normalized by class size, while numbers show absolute values). As expected, IBS performs worse on non-synthetic data and the misclassification rate for both SC and IBS is substantially higher than for* RAID.

between some relationships (e.g., surrounded versus surrounding). The RAID descriptor captures the *spatial distribution* of simple relationships over a region, resulting in a more discriminative classifier.

Classification Performance on the Web Dataset. The web dataset consists of 69 images containing a total of 121 manually labeled relationships. These relationships represent a reasonably balanced mix of the complex relationship classes shown in Figure 4. Since good examples of the 'leaning' relationship are quite uncommon, we used the 'rising' relationship ('hanging' mirrored horizontally) instead. Similar to the synthetic dataset, we used one binary k-NN classifier with $k = 5$ for each relationship class to perform the classification.

Results of a leave-one-out cross-validation and a comparison to IBS and Shape Contexts are shown in Figure 9. The results for Shape Contexts are similar to those of the synthetic dataset. Some classes like 'hanging' and 'arching over' cannot be detected and many relationships were incorrectly classified as not belonging to any class (last column of the confusion matrix). Since IBS is sensitive to noise and geometric variations of the interacting surfaces, it performs worse on this non-synthetic dataset, with a high misclassification rate. Our RAID descriptor achieves roughly a 40% increase in the F_1 score compared to Shape Contexts, 55% compared to IBS, and can successfully classify most of the regions.

Automatically Segmented Regions. Recent methods for semantic image segmentation [Zheng et al. 2015; Long et al. 2015; Chen et al. 2015] can find and label image regions with sufficient quality to use as input for methods that analyze higher-level properties of images. We use the method by Zheng et al. [2015] to demonstrate the performance of our descriptor on automatically segmented regions. Currently, 20 types of objects can be detected, including persons, horses, bottles and motorbikes. To provide a fair comparison with manual labeling, we only use queries in which the subjects and objects most frequently encountered in the queried relationship are part of these 20 object types. This is the case in the 'riding,' 'leaning,' 'crossing,' and 'holding' relationships. Results are shown in Figure 10. As expected, the performance is lower than for manually segmented regions, primarily due to incorrectly merged regions and misclassifications. It is, however, clearly above the performance of both Shape Contexts and IBS, even when applied to manually segmented regions. Using automatic segmentation, our descriptor can be applied directly to image databases without the need for manually annotated regions.

Noisy Data. To evaluate the robustness of our descriptor to noisy image regions, we created several noisy versions of our datasets. Boundary noise was added to the image regions by re-sampling the boundary and adding a normally distributed offset to all vertices with variance of s times the largest image dimension. We varied s from 0.005 to 0.379, nearly two orders of magnitude. Classification and retrieval performance of our descriptor on the noisy datasets are shown in Figure 11. The performance is relatively stable up to a noise strength of approximately 0.07, where the performance be-

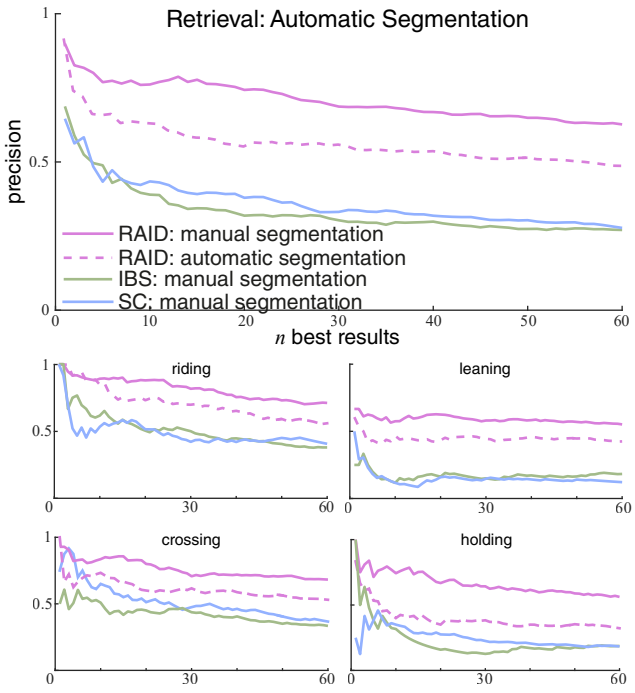


Figure 10: *Retrieval performance on automatically segmented image regions, using the method of Zheng et al. [2015]. On the top, we show the average performance over the four queries shown at the bottom.* RAID*'s performance is better than the performance IBS and Shape Contexts when applied to* manually *segmented regions in all four queries.*

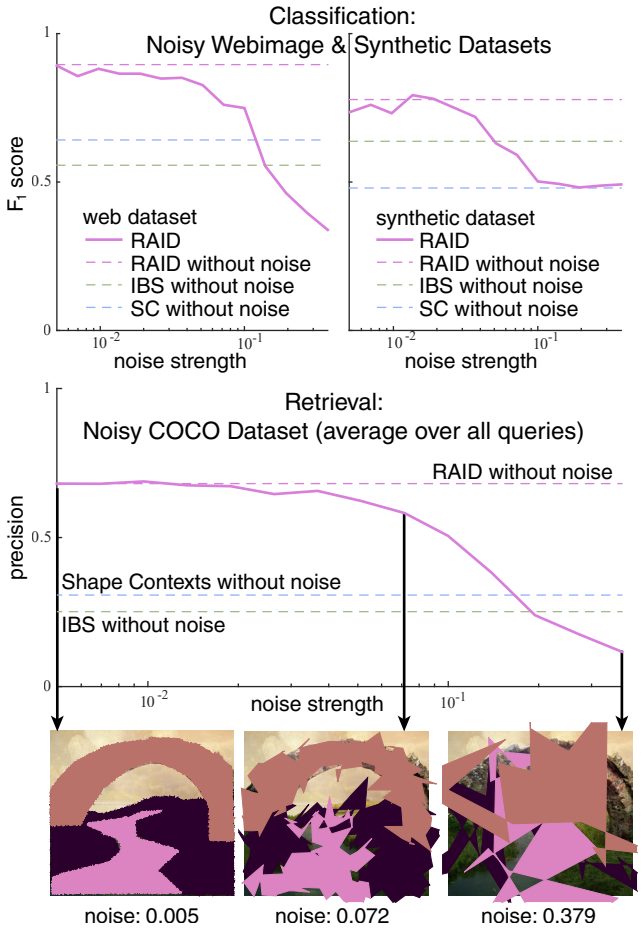


Figure 11: *Performance on datasets with different levels of noise. On the top, we show classification performance on both the web- and the synthetic datasets over increasing levels of noise. In the center, we show retrieval performance for the first* 20 *images returned from the COCO dataset. Performance with manually segmented regions is shown as dotted lines for reference. Examples of regions at different noise levels are shown at the bottom.* RAID*'s performance is relatively stable for a wide range of noise levels.*

gins to drop off. Note that at these noise levels, the regions become too distorted to be discernible even for a human observer. Note that this robustness quality also ensures we can use RAID for relation-driven, sketch-based retrieval (see also supplementary video).

Limitations. Since we do not have depth data available, our descriptor is limited to relations between 2D regions, not 3D objects. In future work we would like to explore possibilities to extend our descriptor to annotated RGBD images.

Due to the limited number of bins of our descriptor (256 in our experiments), there is a limit to the complexity of the relationships that can be described. An 'interleaved' relationship, for example, might be difficult to describe. Figure 12 presents an example. Here, the interleaved rings of the center regions cannot be distinguished properly from rings of the query, since the detail is too fine to be captured by the descriptor bins. Increasing the resolution of the descriptor relieves the problem but also makes the descriptor less tolerant to geometric differences in the relationships. In future work, we would like to experiment with different distance measures, such as the Earth-Movers distance [Rubner et al. 1998], which might help to increase the resolution of the descriptor without decreasing the tolerance.

7 Conclusion

We have presented RAID, a descriptor for complex relationships between image regions. The key idea of the descriptor is to capture the spatial distribution of simple point-to-region relationship to describe more complex relationships between a pair of regions. To the best of our knowledge, there is currently no descriptor that attempts to capture complex relationships between image regions. Our descriptor is conceptually simple, easy to implement and experiments have shown that it can be employed successfully for relationship-based image retrieval in large databases and for relationship classification, with a clear advantage over Shape Contexts, a descriptor for simple point-to-region relationships, and IBS, a state-of-the-art descriptor for shape relationships.

Continuing this line of research, we would like to extend RAID to describe relationships between 3D models (either given as voxels or polygon meshes), use our descriptor in more advanced machine learning techniques, for instance to refine a query by interactively marking good and bad results, and use RAID as a basis to describe the composition of an image, for example by constructing a graph of pair-wise region relationships.

Acknowledgements

We thank the participants of our user study, the anonymous reviewers for their comments and constructive suggestions and Shuai Zheng for giving us early access to the semantic segmentation code. The research described here was supported by the Office of Sponsored Research (OSR) under Award No. OCRF-2014-CGR3-62140401, the Visual Computing Center at KAUST, ERC Starting Grant SmartGeometry (StG-2013-335373), Marie Curie CIG 303541 and the Open3D Project (EPSRC Grant EP/M013685/1).

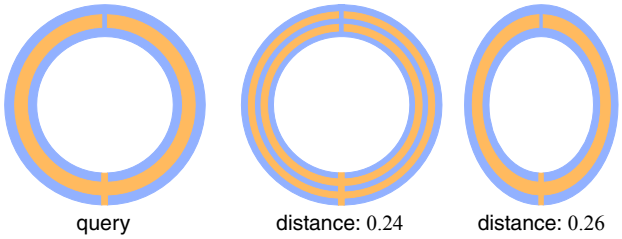


Figure 12: *The resolution of our descriptor limits the complexity of relationships that can be captured. The query (left) is more similar to the image in the center than to a deformed version of the query (right), since details in the center are too fine to be described properly. Shown are the L_1 distances of the descriptors (maximum possible distance is 2).*

References

ARNOLD, S., M., W., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content-based image retrieval at the end of the early years. *IEEE PAMI 22*, 12 (Dec.), 1349–1380.

BADADAPURE, P. R. 2013. Content-Based Image Retrieval by Combining Structural and Content Based Features. *International Journal of Engineering and Advanced Technology 2*, 4, 154–156.

BELONGIE, S., MALIK, J., AND PUZICHA, J. 2002. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 24*, 4, 509–522.

BERTHOUZOZ, F., LI, W., DONTCHEVA, M., AND AGRAWALA, M. 2011. A framework for content-adaptive photo manipulation macros: Application to face, landscape, and global manipulations. *ACM TOG 30*, 5 (Oct.), 120:1–120:14.

BLOCH, I. 2005. Fuzzy spatial relationships for image processing and interpretation: A review. In *Image and Vision Computing*, vol. 23, 89–110.

2015. Boost polygon, version 1.58. www.boost.org.

CAO, Y., WANG, C., ZHANG, L., AND ZHANG, L. 2011. Edgel index for large-scale sketch-based image search. In *IEEE CVPR*, 761–768.

CELEBI, M. E., AND ASLANDOGAN, Y. A. 2005. A comparative study of three moment-based shape descriptors. In *IEEE Proc. of the Internat. Conf. on Information Technology*, 788–793.

CHANDRAN, S., AND KIRAN, N. 2003. Image retrieval with embedded region relationships. In *Proceedings of SAC*, 760.

CHAO, Y.-W., WANG, Z., HE, Y., WANG, J., AND DENG, J. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*.

CHEN, T., CHENG, M.-M., TAN, P., SHAMIR, A., AND HU, S.-M. 2009. Sketch2photo: Internet image montage. *ACM TOG 28*, 5 (Dec.), 124:1–124:10.

CHEN, K., LAI, Y.-K., WU, Y.-X., MARTIN, R., AND HU, S.-M. 2014. Automatic semantic modeling of indoor scenes from low-quality rgb-d data using contextual information. *ACM TOG 33*, 6 (Nov.), 208:1–208:12.

CHEN, L., PAPANDREOU, G., KOKKINOS, I., MURPHY, K., AND YUILLE, A. L. 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. *ICLR* (Nov.).

CHOI, W., SHAHID, K., AND SAVARESE, S. 2009. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops*, 1282–1289.

CHUA, T. S., TAN, K.-L., AND OOI, B. C. 1997. Fast signature-based color-spatial image retrieval. In *Multimedia Computing and Systems '97. Proceedings., IEEE International Conference on*, 362–369.

EITZ, M., HILDEBRAND, K., BOUBEKEUR, T., AND ALEXA, M. 2009. A descriptor for large scale image retrieval based on sketched feature lines. In *Eurographics Symposium on Sketch-Based Interfaces and Modeling*, 29–38.

EITZ, M., HILDEBRAND, K., BOUBEKEUR, T., AND ALEXA, M. 2009. A descriptor for large scale image retrieval based on sketched feature lines. In *SBIM '09*, ACM, New York, NY, USA, 29–36.

EITZ, M., RICHTER, R., HILDEBRAND, K., BOUBEKEUR, T., AND ALEXA, M. 2011. Photosketcher: Interactive sketch-based image synthesis. *Computer Graphics and Applications, IEEE 31*, 6 (Nov), 56–66.

EITZ, M., RICHTER, R., BOUBEKEUR, T., HILDEBRAND, K., AND ALEXA, M. 2012. Sketch-based shape retrieval. *ACM TOG 31*, 4 (July), 31:1–31:10.

FISHER, M., SAVVA, M., AND HANRAHAN, P. 2011. Characterizing structural relationships in scenes using graph kernels. In *ACM TOG*, vol. 30, ACM, 34.

FISHER, M., RITCHIE, D., SAVVA, M., FUNKHOUSER, T., AND HANRAHAN, P. 2012. Example-based synthesis of 3d object arrangements. In *ACM SIGGRAPH Asia*.

FISHER, M., SAVVA, M., LI, Y., HANRAHAN, P., AND NIESSNER, M. 2015. Activity-centric scene synthesis for functional 3d scene modeling. *ACM TOG 34*, 6.

FLUSSER, J. 1992. Invariant shape description and measure of object similarity. In *Image Processing and its Applications, 1992., International Conference on*, 139–142.

GOSHTASBY, A. 1985. Description and discrimination of planar shapes using shape matrices. *IEEE PAMI 7*, 6, 738–743.

HAYS, J., AND EFROS, A. A. 2007. Scene completion using millions of photographs. *ACM TOG 26*, 3 (July).

HSIEH, S.-M., AND HSU, C.-C. 2008. Retrieval of images by spatial and object similarities. *Inf. Process. Manage. 44*, 3 (May), 1214–1233.

HU, S.-M., ZHANG, F.-L., WANG, M., MARTIN, R. R., AND WANG, J. 2013. PatchNet: A Patch-based Image Representation for Interactive Library-driven Image Editing. *ACM TOG 32*, 6, 1–12.

HU, R., ZHU, C., VAN KAICK, O., LIU, L., SHAMIR, A., AND ZHANG, H. 2015. Interaction context (icon): Towards a geometric functionality descriptor. *ACM TOG 34*, 4 (July), 83:1–83:12.

HUANG, H., YIN, K., GONG, M., LISCHINSKI, D., COHEN-OR, D., ASCHER, U., AND CHEN, B. 2013. "mind the gap": Tele-registration for structure-driven image completion. *ACM TOG 32*, 6 (Nov.), 174:1–174:10.

HUANG, S., WANG, W., AND ZHANG, H. 2014. Retrieving images using saliency detection and graph matching. In *IEEE ICIP*, 3087–3091.

JANSEN, S., SHANTIA, A., AND WIERING, M. A. 2015. The neural-sift feature descriptor for visual vocabulary object recognition. In *IJCNN*, 1–8.

KARPATHY, A., AND LI, F.-F. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE CVPR*.

KAZMI, I. K., YOU, L., AND ZHANG, J. J. 2013. A survey of 2d and 3d shape descriptors. *2014 11th International Conference on Computer Graphics, Imaging and Visualization 0*, 1–10.

KIM, V. G., CHAUDHURI, S., GUIBAS, L., AND FUNKHOUSER, T. 2014. Shape2Pose: Human-Centric Shape Analysis. *ACM SIGGRAPH 33*, 4.

KO, B., AND BYUN, H. 2002. Multiple Regions and Their Spatial Relationship-Based Image Retrieval. In *LNCS 2383*. 81–90.

KRISHNA, R., ZHU, Y., GROTH, O., JOHNSON, J., HATA, K., KRAVITZ, J., CHEN, S., KALANTIDIS, Y., LI, L.-J., SHAMMA, D. A., BERNSTEIN, M., AND FEI-FEI, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

KULKARNI, G., PREMRAJ, V., ORDONEZ, V., DHAR, S., LI, S., CHOI, Y., BERG, A. C., AND BERG, T. L. 2013. Baby talk: Understanding and generating simple image descriptions. *IEEE PAMI 35*, 12, 2891–2903.

LAN, T., YANG, W., WANG, Y., AND MORI, G. 2012. Image retrieval with structured object queries using latent ranking SVM. In *Lect. Notes in Computer Science*, vol. 7577 LNCS, 129–142.

LAN, T., RAPTIS, M., SIGAL, L., AND MORI, G. 2013. From subcategories to visual composites: A multi-level framework for object detection. In *IEEE ICCV*.

LEE, S. L. S., AND HWANG, E. H. E. 2002. Spatial similarity and annotation-based image retrieval system. *Proceedings of Fourth Int. Symposium on Multimedia Software Engineering*.

LIN, T., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., AND ZITNICK, C. L. 2014. Microsoft COCO: common objects in context. *CoRR abs/1405.0312*.

LIU, T., CHAUDHURI, S., KIM, V. G., HUANG, Q.-X., MITRA, N. J., AND FUNKHOUSER, T. 2014. Creating Consistent Scene Graphs Using a Probabilistic Grammar. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia) 33*, 6.

LONG, J., SHELHAMER, E., AND DARRELL, T. 2015. Fully convolutional networks for semantic segmentation. *IEEE CVPR*.

LOWE, D. 2004. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision 60*, 2, 91–110.

MALISIEWICZ, T., AND A., E. A. 2009. Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships. In *NIPS*, 1–9.

OOI, B. C., TAN, K.-L., CHUA, T. S., AND HSU, W. 1998. Fast image retrieval using color-spatial information. *The VLDB Journal 7*, 2, 115–128.

PENTLAND, A., PICARD, R. W., AND SCLAROFF, S. 1996. Photobook: Content-based manipulation of image databases. *Int. J. Comput. Vision 18*, 3 (June), 233–254.

RUBNER, Y., TOMASI, C., AND GUIBAS, L. J. 1998. A metric for distributions with applications to image databases. IEEE Computer Society, Washington, DC, USA, IEEE ICCV, 59–66.

SADEGHI, M. A., AND FARHADI, A. 2011. Recognition using visual phrases. IEEE Computer Society, Washington, DC, USA, IEEE CVPR, 1745–1752.

SHAO, T., MONSZPART, A., ZHENG, Y., KOO, B., XU, W., ZHOU, K., AND MITRA, N. 2014. Imagining the unseen: Stability-based cuboid arrangements for scene understanding. *ACM SIGGRAPH Asia*. * Joint first authors.

SHECHTMAN, E., AND IRANI, M. 2007. Matching local self-similarities across images and videos. In *IEEE CVPR*, 1–8.

SMITH, J. R., AND CHANG, S.-F. 1996. Visualseek: A fully automated content-based image query system. In *Proceedings of the Fourth ACM International Conference on Multimedia*, ACM, New York, NY, USA, MULTIMEDIA '96, 87–98.

TEAGUE, M. R. 1980. Image analysis via the general theory of moments∗. *J. Opt. Soc. Am. 70*, 8 (Aug), 920–930.

WANG, J., AND HUA, X.-S. 2011. Interactive image search by color map. *ACM Trans. Intell. Syst. Technol. 3*, 1, 12:1–12:23.

WANG, Y.-H., 2003. Image indexing and similarity retrieval based on spatial relationship model.

XU, H., WANG, J., HUA, X.-S., AND LI, S. 2010. Image search by concept map. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, USA, SIGIR '10, 275–282.

XU, K., CHEN, K., FU, H., SUN, W.-L., AND HU, S.-M. 2013. Sketch2scene: Sketch-based co-retrieval and co-placement of 3d models. *ACM TOG 32*, 4 (July), 123:1–123:15.

YÜCER, K., JACOBSON, A., HORNUNG, A., AND SORKINE, O. 2012. Transfusive image manipulation. *ACM TOG 31*, 6 (Nov.), 176:1–176:9.

ZHANG, D., AND LU, G. 2004. Review of shape representation and description techniques. *Pattern Recognition 37*, 1, 1 – 19.

ZHAO, X., WANG, H., AND KOMURA, T. 2014. Indexing 3d scenes using the interaction bisector surface. *ACM TOG 33*, 3 (June), 22:1–22:14.

ZHENG, Y., COHEN-OR, D., AVERKIOU, M., AND MITRA, N. J. 2014. Recurring part arrangements in shape collections. *Computer Graphics Forum*.

ZHENG, S., JAYASUMANA, S., ROMERA-PAREDES, B., VINEET, V., SU, Z., DU, D., HUANG, C., AND TORR, P. 2015. Conditional random fields as recurrent neural networks. In *IEEE ICCV*.

ZHOU, X. M., ANG, C. H., AND LING, T. W. 2001. Image retrieval based on object's orientation spatial relationship. *Pattern Recognition Letters 22*, 5, 469–477.