

# RAID: A Relation-Augmented Image Descriptor

Paul Guerrero\*

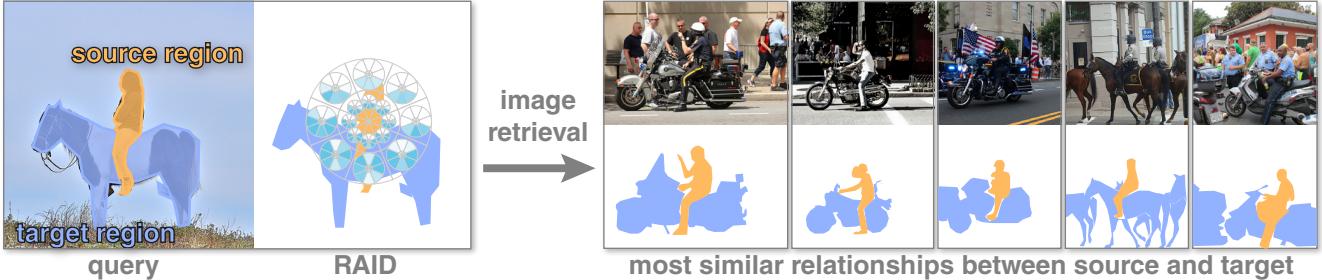
KAUST, University College London

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 image retrieval with queries based on complex relationships between regions, such as the ‘riding’ relationship between the orange source and the blue target region.

## Abstract

As humans, we regularly interpret images based on the relations between image regions. For example, a person *riding* object X, or a plank *bridging* two objects. Current methods provide limited support to search for images 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 capture the spatial distribution of simple point-to-region relationships to describe more complex relationships between two image regions. We evaluate the proposed descriptor by querying into a large subset of the Microsoft COCO database and successfully extract non-trivial images demonstrating complex inter-region relations, which are easily missed or erroneously classified by existing methods.

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

## 1 Introduction

Content-based image retrieval is an important task for image processing applications. For example, an artist may search for a particular scene configuration for inspiration, or a media creator might seek images with a particular assembly of objects. Text-based search, using keywords or tags, is still the most commonly available search option. Advanced alternatives exist that exploit color histograms [Pentland et al. 1996; Smeulders et al. 2000], object sketches [Mathias Eitz and Alexa 2009; Cao et al. 2011], or even a rough composition guidance [Hu et al. 2013].

The last decades have witnessed significant advances in semi-automated and automated image segmentation algorithms. They have resulted in large image databases containing many thousands of labeled and segmented images (e.g., Microsoft COCO [Lin et al. 2014], Pascal VOC [Mottaghi et al. 2014], MIT SUN [Xiao et al. 2010]). Hence, it is now possible to search for images having regions labeled ‘horse,’ or ‘man,’ or both ‘horse’ *and* ‘man.’ However, there is little support to query based on how the segments are *related*. For example, how can we search for images showing ‘man *riding* a horse,’ or ‘man *standing next to* a horse,’ or more generally ‘man *riding* any object.’

In this paper, we present RAID as a relation-augmented image descriptor that supports queries based on inter-segment relations. We identify a set of commonly occurring relations, particularly complex relations (e.g., bridging, riding, leaning, etc.) beyond usual relations like above, below, adjacent, etc. This essentially allows us to query by verbs relating image segment names by associating a particular descriptor with each of the verbs. Our framework is general in the sense that the user can alternatively sketch a composition of image segments, or pick a pair of regions in an existing image, and the system can construct an appropriate RAID. An example is given in Figure 1, where the riding relationship in an existing image is used to query a large annotated image database. These relationships describe the spatial composition of regions in an image. It is important to note that we are tackling a purely two-dimensional problem. Describing the relationship between the actual three-dimensional objects that are represented by the regions is a different problem.

Inter-region relationship are useful in several active research areas, including image editing, image synthesis and content-based image retrieval. They could 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, and enhance image completion [Hays and Efros 2007; Huang et al. 2013] in context-dependent image regions.

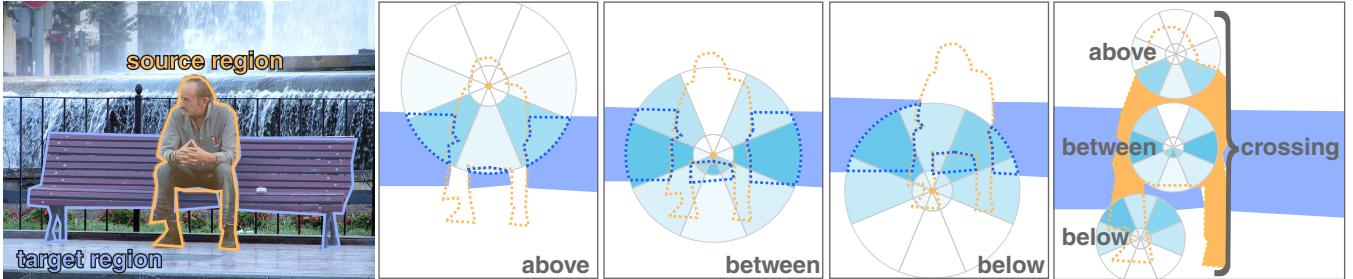
Current shape descriptors such as Shape Contexts [Belongie et al. 2002] are able to describe the relationship between a point and a region, such as ‘below’ or ‘adjacent.’ In a complex relationship between two regions, these simple point-to-region relationships usually vary over a region. Take for example the image in Figure 2: the head of the man is above the bench, while his feet are below. The key idea of our descriptor is to capture the spatial distribution of these simple point-to-region relationships to describe more complex relationships between two image regions.

We evaluate the query performance as well as the classification performance of our descriptor and provide a comparison to Shape Contexts as a baseline shape descriptor. The query performance is measured quantitatively 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. Results show that our method is able to successfully describe complex relationships with a clear improvement over Shape Contexts.

\* paul.guerrero@ucl.ac.uk

† n.mittra@cs.ucl.ac.uk

‡ pwonka@gmail.com



**Figure 2:** 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 is 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.

## 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 individual 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. In these methods, no attempt is made to describe complex relationships or capture a spatial distribution of relationships. More recent approaches can be classified by the type of models they employ:

**Point-based relationships 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 in 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 point.

**String-based relationships 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’).

**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 2-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.

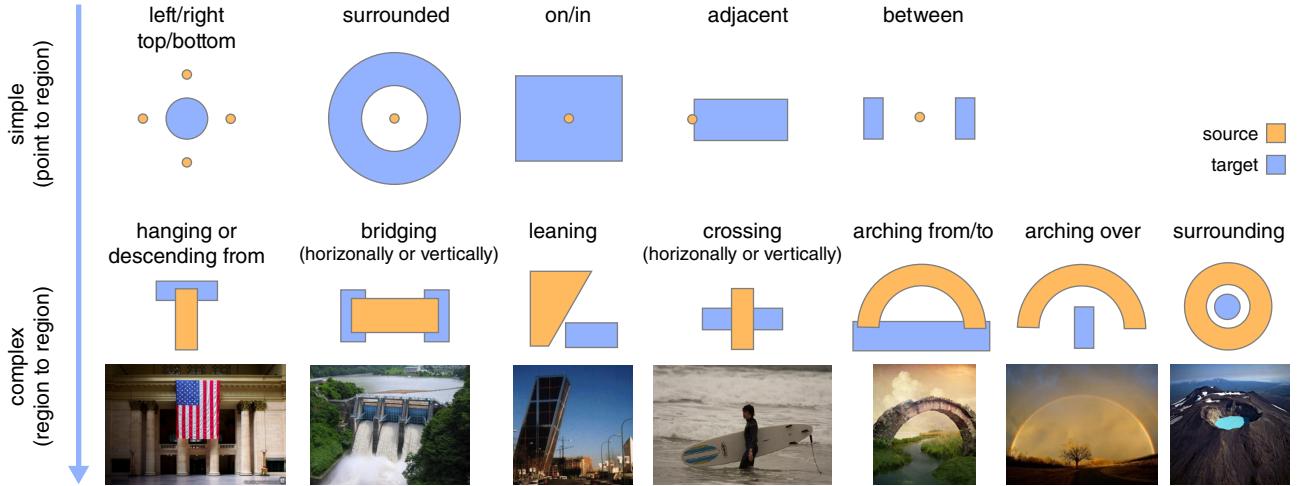
**Scene understanding.** An important part of scene understanding is to accurately identify the relationship between scene objects. Several methods tackle this challenge by creating models of region relationships. Malisiewicz and Efros [2009] encode the spatial context of image regions in a graph. Features 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 like ‘above’, ‘on’, and ‘near’. Adding an additional class requires implementing an additional detector. Recently, Karpathy et al. [2015] presented a deep learning method to create natural language descriptions of images, with impressive results when trained on large datasets. However, it does not have an explicit representation of relationships. Our approach describes more complex relationships, provides a single data-driven descriptor for all relationship classes and does not need to be trained on a large dataset.

**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 to compare the performance of our method.

## 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 3. We can observe, that most of



**Figure 3:** Classes of spatial relationships between two-dimensional image regions. We distinguish simple relationships (top row) and complex relationships (bottom row) between the orange source region and the blue target regions. Example images are shown below each complex relationship.

the relationships are asymmetrical. For example, if a region  $A$  is to the left of region  $B$ , then region  $B$  is to the right of region  $A$ . We therefore need to distinguish between the two regions involved in a relationship and 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 regions. For example, both a point and a region can be surrounded by 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. Hence, the ‘surrounded’ relationship is simple and the ‘surrounding’ relationship complex. In Figure 3 examples of simple relationships are shown on top and examples of complex relationship are shown on the bottom.

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 the individual point to region descriptors. In the following we will 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 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]:

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

where  $\Phi_i$  and  $R_j$  are 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*. Figure 4, center shows an example for 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 corresponding to the fraction of its area covered by region  $T$ .

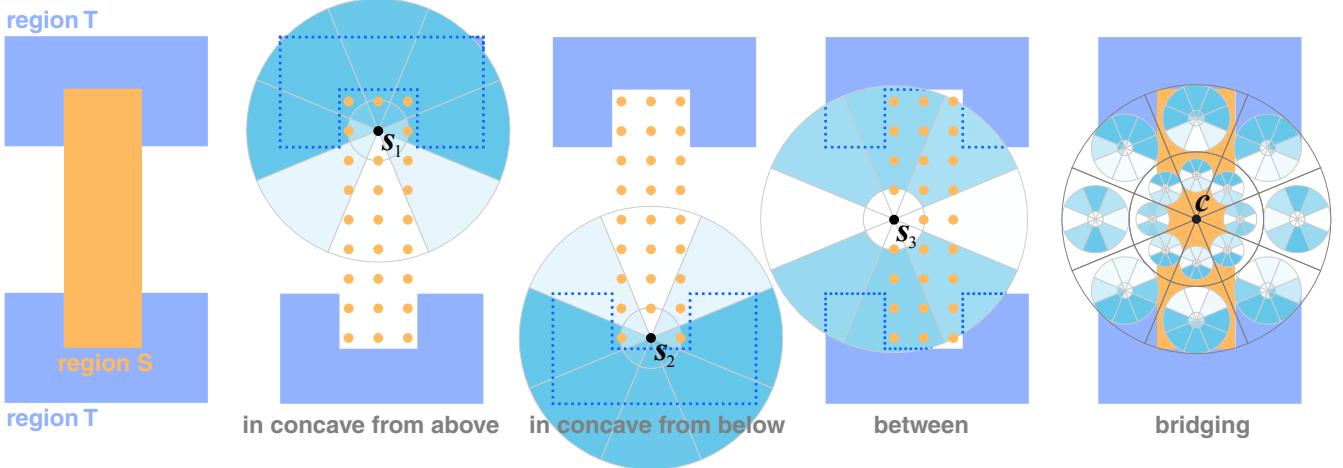
The distribution of point relationships over the source region is then encoded by a second histogram  $\mathcal{H}^S$  over the individual point histograms, resulting in a four-dimensional histogram:

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

where  $\mathbf{c}$  is the centroid of the source region. Figure 4, right shows an illustration of the 4D histogram. The denominator normalizes each bin by the intersection of the bin area with the source region. This factors out the dependence of the histogram on the exact shape of the source region and only captures the distribution of point histograms. Bins with zero intersection are assigned the value of the point histogram at the closest point of the source region. Finally, we perform a histogram normalization:

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

We call this histogram the RAID descriptor. Similar to the SIFT descriptor [Lowe 2004], the RAID descriptor is a histogram of his-



**Figure 4:** 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  $s$  in  $S$  and the region  $T$  are described by histograms of relative distance and direction from  $s$  to points in  $T$ :  $s_1$  and  $s_2$  are in a concave of  $T$ , while  $s_3$  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  $s_3$  that are between  $T$  are added to bins closer to the centroid  $c$ , while points like  $s_1$  and  $s_2$  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.

tograms, but RAID encodes directions and distances to a target region while SIFT encodes gradient orientations.

## 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. A performant 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, therefore, we resort to an approximation. First, point histograms are computed at a regular grid of samples  $s$  inside the source region. As a good tradeoff between performance and accuracy, the density is chosen to be approximately  $10000/a_1$ , where  $a_1$  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:

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

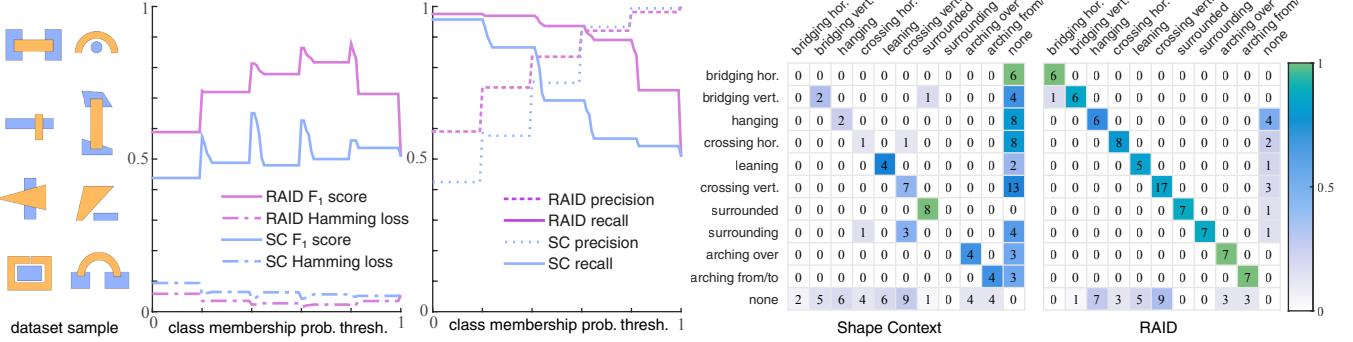
where  $S$  is the set of samples inside the source region,  $\mathbf{b}_{kl}$  is the centroid of bin  $(k, l)$  and  $\mathcal{G}(\mathbf{x}|\boldsymbol{\mu}, \sigma^2)$  is an isotropic two-dimensional Gaussian with mean  $\boldsymbol{\mu}$  and variance  $\sigma^2$ . The variance of the Gaussian 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:

$$\mathcal{H}_{ijkl}^S = \frac{\hat{\mathcal{H}}_{ijkl}^S}{\sum_{ijkl} \hat{\mathcal{H}}_{ijkl}^S}. \quad (5)$$

In all our experiments, we set the maximum distance  $r_{\max}$  for the outermost bin in the RAID descriptor to the maximum distance between 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 4. Center images 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’), therefore we 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. However, this would result in severe undersampling of the more uncommon relationship classes. Considering this, we collected a set of 69 images from the web instead, containing a balanced mix of relationship classes. All datasets were finalized before starting our



**Figure 5:** Classification performance on the synthetic dataset and comparison to 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. On the right, the confusion matrices for Shape Contexts and RAID are shown (rows correspond to actual classes, columns to predicted classes, colors are normalized by class size, while numbers show absolute values). Note that Shape Contexts generally perform worse and are unable to detect some relationship classes, like ‘bridging’ or ‘surrounding’.

experiments.

To the best of our knowledge, currently there exists no descriptor that explicitly attempts to describe complex relationships between image regions. Most methods only describe simple relationships, that is, 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]. 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 of 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 pre-computed descriptors, which requires roughly 0.46 seconds in our Matlab implementation.

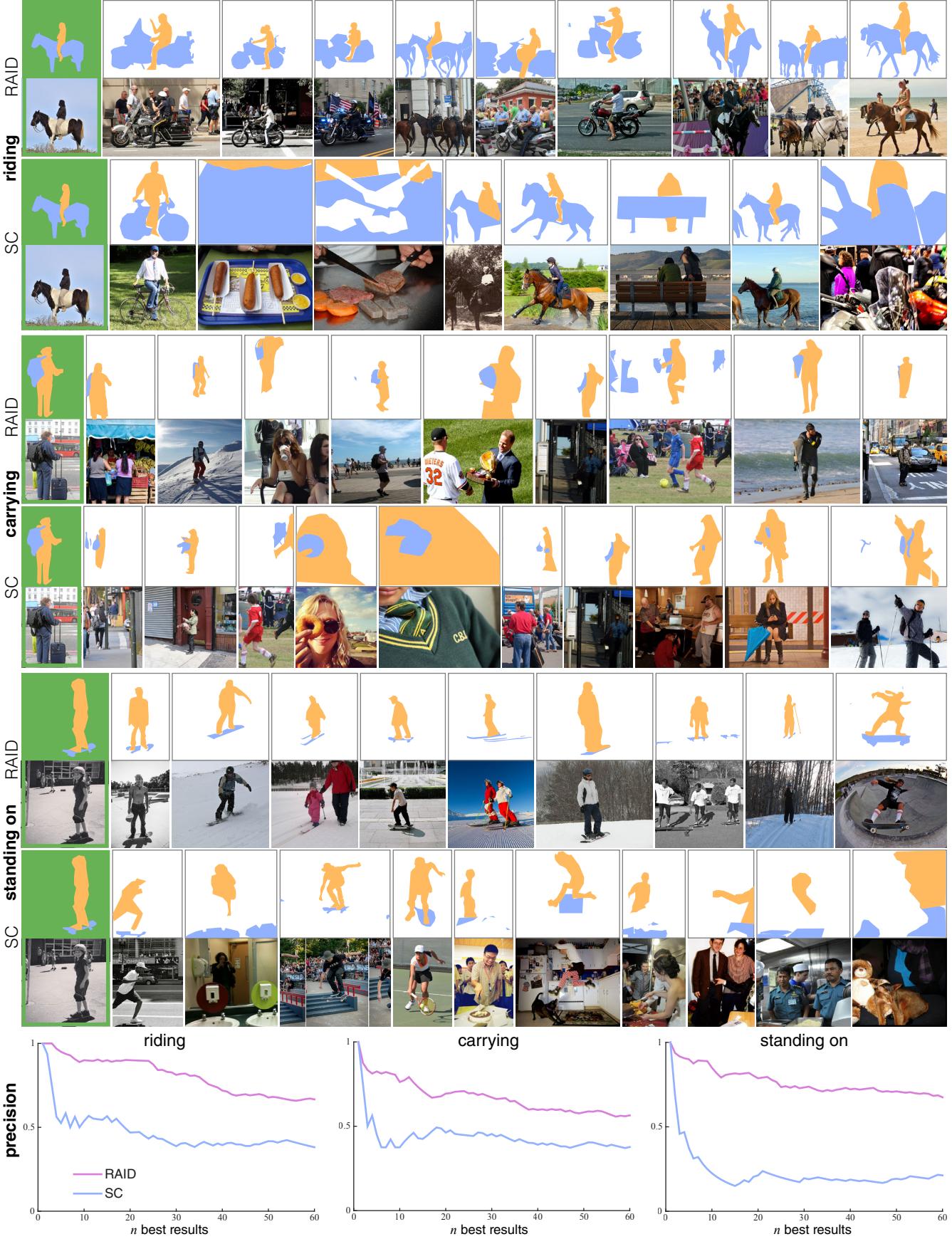
**Image Retrieval** An interesting area of application for the RAID descriptor is image retrieval from large databases. Our method can extend the search capability of a system by enabling queries for given relationships, such as ‘riding’ or ‘standing on’. In the following we present experiments we have 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 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.

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  $n$ . The ground truth was created by three persons who manually annotated the results of the query in randomized order, without knowledge of our descriptor and without knowledge which method generated the results. In the ‘riding’ query (Figure 6, first row), the source region contains an interesting distribution of simple relationship, 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 also return many false positives with 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.

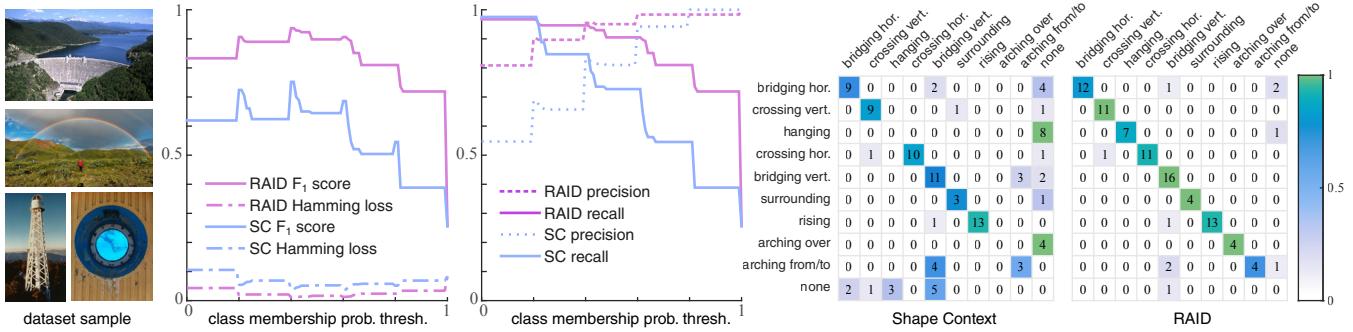
**Classification Performance on the Synthetic Dataset** We performed additional evaluation on a small synthetic dataset containing



**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 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 web image dataset and comparison to 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 for Shape Contexts and RAID are shown (rows correspond to actual classes, columns to predicted classes, colors are normalized by class size, while numbers show absolute values). Similar to the synthetic dataset, Shape Contexts have a lower performance and have problems detecting some of the classes.

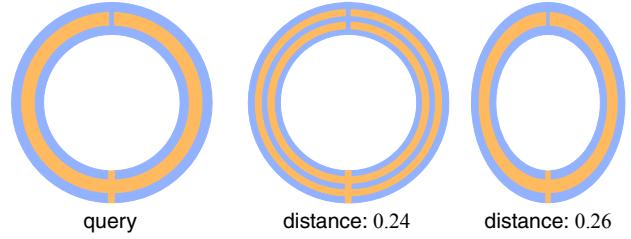
164 manually created images. Each image shows a single source- and a single target region. These region pairs were labeled manually with zero, one, or multiple labels from among the seven complex relationship classes shown in Figure 3 plus the ‘surrounded’ relationship. Of the 164 relationships, 97 are 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 Shape Contexts are shown in Figure 5. Since Shape Contexts only capture simple relationships between a point and a region, they perform poorly for 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. The RAID descriptor captures the *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 3. Since good examples of a ‘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 Shape Contexts are shown in Figure 8. Note how the results 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). Our RAID descriptor achieves roughly a 40% increase in the  $F_1$  score compared to Shape Contexts and can successfully classify most of the regions.

**Limitations** 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,



**Figure 9:** 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 RAIDs (maximum possible distance is 2).

for example, might be difficult to describe. See Figure 9 for 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.

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.

## References

- 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 Trans. Graph.* 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](http://www.boost.org).
- CAO, Y., WANG, C., ZHANG, L., AND ZHANG, L. 2011. Edgel index for large-scale sketch-based image search. In *Proc. of IEEE Conf. on Comp. Vision and Pattern Recognition*, 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 the 2003 ACM symposium on Applied computing - SAC '03*, 760.
- 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 Trans. Graph.* 26, 3 (July).
- HSIEH, S. M., AND HSU, C. C. 2008. Retrieval of images by spatial and object similarities. *Information Processing and Management* 44, 3, 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 Transactions on Graphics* 32, 6, 1–12.
- 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 Trans. Graph.* 32, 6 (Nov.), 174:1–174:10.
- HUANG, S., WANG, W., AND ZHANG, H. 2014. Retrieving images using saliency detection and graph matching. In *2014 IEEE Int. Conference on Image Processing (ICIP)*, 3087–3091.
- KARPATHY, A., AND LI, F.-F. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *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.
- KO, B., AND BYUN, H. 2002. Multiple Regions and Their Spatial Relationship-Based Image Retrieval. In *LNCS* 2383. 81–90.
- 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 Trans. on Pattern Anal. and Mach. Intell.* 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.
- 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*.
- LOWE, D. 2004. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision* 60, 2, 91–110.
- MALISIEWICZ, T., AND EFROS, A. A. 2009. Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships. In *NIPS*, 1–9.
- MATHIAS EITZ, KRISTIAN HILDEBRAND, T. B., 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.
- MOTTAGHI, R., CHEN, X., LIU, X., CHO, N.-G., LEE, S.-W., FIDLER, S., URTASUN, R., AND YUILLE, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *IEEE CVPR*.
- 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. In *Proc. of the Sixth International Conference on Computer Vision*, IEEE Computer Society, Washington, DC, USA, ICCV '98, 59–66.
- SMEULDERS, A. W. M., WORRING, M., SANTINI, S., GUPTA, A., AND JAIN, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 12 (Dec.), 1349–1380.
- TEAGUE, M. R. 1980. Image analysis via the general theory of moments\*. *J. Opt. Soc. Am.* 70, 8 (Aug), 920–930.
- WANG, Y.-H., 2003. Image indexing and similarity retrieval based on spatial relationship model.
- XIAO, J., HAYS, J., EHINGER, K., OLIVA, A., AND TORRALBA, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE CVPR*.
- YÜCER, K., JACOBSON, A., HORNUNG, A., AND SORKINE, O. 2012. Transfusive image manipulation. *ACM Trans. Graph.* 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.
- 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.