

A BAG-OF-REGIONS APPROACH TO SKETCH-BASED IMAGE RETRIEVAL

Rui Hu, Tinghuai Wang, and John Collomosse

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, Surrey, UK.

ABSTRACT

This paper presents a system for retrieving photographs using free-hand sketched queries. Regions are extracted from each image by gathering nodes of a hierarchical image segmentation into a bag-of-regions (BoR) representation. The BoR represents object shape at multiple scales, encoding shape even in the presence of adjacent clutter. We extract a shape representation from each region, using the Gradient Field HoG (GF-HOG) descriptor which enables direct comparison with the sketched query. The retrieval pipeline yields significant performance improvements over the previous GF-HOG results reliant on single-scale Canny edge maps, and over leading descriptors (SIFT, SSIM) for visual search. In addition, our system enables localization of the sketched object within matching images.

Index Terms— Sketch based Image Retrieval (SBIR), Bag-of-Regions, Bag-of-visual-words, GF-HOG.

1. INTRODUCTION

Although text keywords remain a popular method for searching image collections, text is unable to concisely convey the desired appearance of an image. Querying by visual example (QVE) provides an attractive alternative to keyword search for appearance based search. Successes with bag-of-visual-words (BoVW) approaches have led to photo-realistic QVE systems exhibiting leading performance in benchmarking trials (e.g. ImageCLEF, PASCAL) and scalability over large image collections. More recently BoVW has also been proposed for use in sketch based image retrieval (SBIR) systems using free-hand sketched queries depicting desired shape [1] [2]. However, both systems these propose shape descriptors based on low level visual information extracted from photo-realistic images, e.g. dominant edges derived from a single-scale Canny operator, without considering spatial coherence or higher level visual information such as image regions. Canny edges are representative of high frequency information, whilst strong edges are a fundamental cue for SBIR, such edges are often non-trivial to isolate automatically in the presence of clutter or in texture rich images. Recent advances in contour detection and hierarchical segmentation [3] have enhanced object recognition using collections of regions at multiple scales [4]. Such image regions have larger spatial extent and semantic significance than low-level interest point-based features [5]. Regions therefore have appealing properties for object recognition and retrieval, as they naturally encoded both shape and scale information, and provide a

limited domain within which to compute object-level features without being affected by clutter from outside the region.

Inspired by the region based object recognition, this paper proposes to use the bag-of-regions (BoR) approach to SBIR, decomposing images into region representations at multiple scales. Regions are collected from the nodes of a hierarchical region tree generated by [3], ranging in scale from super-pixels to the entire image. We extend the work present in [1] inheriting the Gradient Field HOG (GF-HOG) descriptor and BoVW paradigm first proposed for SBIR. applying this to the bounding contours of the extracted regions at their various scales (Sec. 2). The proposed BoR approach yields significant improvements in performance compared to the results reported in [1] based on three leading descriptors under a classical BoVW framework: GF-HOG[1], SIFT [5], and Self-Similarity (SSIM) [6] (Sec. 3). Matching the sketched shape to the best candidate regions in an image also facilitates localization of the sketched object (subsec. 2.4).

1.1. Related Work

Early sketch based image retrieval systems were typically driven by queries comprising blobs of color or predefined texture [7] [8]. Later systems explored shape descriptors [9] and spectral descriptors such as wavelets [10]. Eitz *et al.* [11] introduced a grid based approach to shape retrieval, dividing the image into regular grids and locate photos using sketched depiction of object shape. Descriptors from each cell were concatenated to form a global image feature. However this offers limited invariance to changes in position, scale or orientation. A depiction invariant descriptor which encapsulates local spatial structure in the sketch and facilitates efficient codebook based retrieval was proposed by Hu *et al.* [1]. This descriptor is able to mitigate the lack of spatial information within a BoVW representation by capturing structure from surrounding regions using a multi-scale HoG descriptor computed over a gradient field interpolated from the orientations of strong Canny edges (GF-HOG). Eitz *et al.* [2] later computed HoG over Canny edges (SHOG) for BoVW though did not interpolate orientations from edges.

There has been some relevant work using regions as the basic elements to address the problem of object retrieval, recognition and segmentation. Sciascio *et al.* [9] investigate extracting shape feature in photo-realistic images using image segmentation. However, in addition to the problem of unstable regions produced by texture-based segmentation algorithm, a particular object can often be either under-segmented or over-segmented, failing to produce an ideal semantically coherent region. feature in photo-realistic im-



Fig. 1. Sample query sketches and some of the top returned images. The localized objects are marked as black regions.

ages using image segmentation. However, in addition to the problem of unstable regions produced by texture-based segmentation algorithm, a particular object can often be either under-segmented or over-segmented, failing to produce an ideal semantically coherent region. Hoiem *et al.* [12] estimate the coarse geometric properties of a single image by learning local appearance and geometric cues on super-pixels even in cluttered natural scenes. Russell *et al.* [13] developed an algorithm that finds and segments visual topics within an unlabeled collection of images by combining multiple candidate segmentations with probabilistic document analysis methods. Todorovic and Ahuja [14] represent objects as region trees and combine structural cues of the trees for matching. Gu *et al.* [4] present a unified framework for object detection, segmentation, and classification using robust overlaid regions produced by a novel region segmentation algorithm [3]. To the best of our knowledge, our approach is the first technique to apply multi-scale region segmentation within a BoVW framework for sketch based image retrieval.

2. SYSTEM OVERVIEW

The pipeline of the system is as follows. First, each image is represented by a bag of regions (BoR) derived from a region tree as shown in Fig. 2. Next, we extract the contours of the region maps containing various levels of detail, and capture local structure in the contour map using the GF-HOG descriptor. Subsequently shape descriptors are clustered via k -means to form a BoVW codebook via k -means and the resulting histogram is used for matching. To facilitate matching, the query sketch is encoded via GF-HOG using the same codebook. Finally, the sketched object is localized in the retrieved image.

2.1. A Bag of Regions

Our system starts by constructing a region tree by performing the hierarchical segmentation algorithm of [3] as shown in Fig. 2. We discard the tree structure and gather all nodes from that tree, except the root which represents the entire image. Intuitively, this bag of regions stores spatially coherent regions at various levels of visual detail. The segmentation hierarchy is obtained by thresholding the Ultrametric Contour Map (UCM) at level 0.4 as described in [3].

Our assumption is that segmentation of the object is approximately correct at some level of this hierarchy (precise segmentation is unnecessary given the ambiguity of sketch and robustness of the GF-HOG descriptor in matching par-

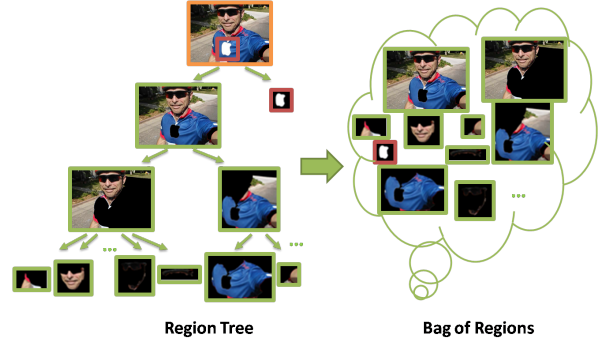


Fig. 2. The “bag of regions” representation of a biker in ETHZ. Regions are collected from all nodes of a region tree generated by [3], ranging in various levels of detail.

tial shape boundaries). Furthermore for the region to be of interest for potential retrieval that object will have distinctive shape that differs noise regions resulting from clutter or segmentation at sub-optimal scale elsewhere in the hierarchy. Such noise regions tends to exhibit similar shapes (highly compact, or highly elongated forms) that due to their number, cluster to form particular visual words analogous to stop words in text retrieval.

2.2. Descriptor

Straightforward approaches to constructing a BoVW codebook and retrieval framework over interest or edge points in an image using local descriptors such as SSIM, SIFT, HoG results in poor retrieval performance as reported in [1]. One explanation is the difficulty of setting a globally appropriate window size for these descriptors, which tend to either capture too little, or integrate too much, of the local edge structure.

Our system builds upon the *gradient field* approach proposed by Hu *et al.* [1]. Their solution is to represent image structure using a dense *gradient field*, interpolated from the sparse set of edges, and then compute local descriptors on the dense *gradient field*. A multi-scale HoG descriptor built on the dense *gradient field*, i.e. GF-HOG is reported to outperform SSIM, SIFT and regular HoG over edges (referred to as EDGE-HOG in [1] and SHOG in [2]). We adopt GF-HOG in our system but evaluation results on SSIM and SIFT descriptors are also given as a baseline in Sec. 3.

Given an edge map $M(x, y) = \{0, 1\}$, a sparse field is computed from the gradient of edge pixels $\theta[x, y] \mapsto \text{atan} \left(\frac{\delta M}{\delta x} / \frac{\delta M}{\delta y} \right)$, $\forall_{x, y} M(x, y) = 1$. Define a dense orientation field Θ_Ω over all coordinates within the region $\Omega \in \mathbb{R}^2$, minimizing:

$$\arg\min_{\Theta} \int_{\Omega} (\nabla \Theta - \mathbf{v})^2 \quad \text{s.t.} \quad \Theta|_{\delta\Omega} = \theta|_{\delta\Omega}. \quad (1)$$

i.e. $\Delta \Theta = 0$ over Ω s.t. $\Theta|_{\delta\Omega} = \theta|_{\delta\Omega}$ for which a discrete solution was presented by Perez *et al.* [15] solving Poisson’s equation with Dirichlet boundary conditions. \mathbf{v} represents the first order derivative of θ .

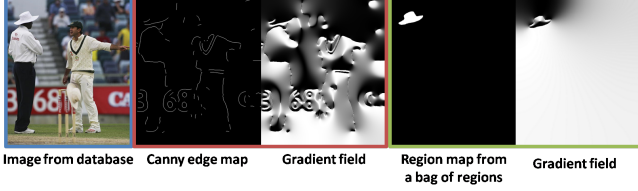


Fig. 3. *Gradient field* of a region map from the bag of regions (green window) and the Canny edge map (red window). The *Gradient field* of the region map captures complete information of the salient shape (the hat) amongst the clutter.

Fig. 3 shows gradient field of a region map from the bag of regions and the Canny edge map. We can see that our “bag of regions” encodes the complete information of salient shapes in the form of enclosed contours of regions present a coherent visual appearance. Moreover, background clutter interferes with region representations only mildly in resulting gradient field.

We compute a HoG descriptor [16] at $\Theta(x, y)$ for all points where $M(x, y) = 1$ in the dense gradient fields of sketch queries and contour maps extracted from the bag of regions to encode the relative location and spatial orientation of edges.

2.3. BoVW Framework for Sketch Based Retrieval

We learn a BoVW codebook by clustering GF-HOG descriptors from the bag of regions of all images via k-means. We construct a frequency histogram H_m^I for m :th contour map I_m of image I and a frequency histogram H^s from the query sketch by quantizing GF-HOG from the sketch using the same codebook. Regions of images are ranked according to histogram similarity $d(H_m^I, H^s)$.

$$d(H^S, H_m^I) = \sum_{i=1}^k \sum_{j=1}^k (\omega_{ij} \min(H^S(i), H_m^I(j))),$$

$$\omega_{ij} = 1 - |\mathcal{H}^S(i) - \mathcal{H}_m^I(j)|. \quad (2)$$

where $H(i)$ indicates the i^{th} bin of the histogram, $\mathcal{H}(i)$ is the normalized visual word corresponding to the i^{th} bin. Initial ranking may list multiple region maps from a single image, and we only select the best ranked region maps of each image to represent the retrieved image.

2.4. Object Localization

As it is desirable to locate the sketched shape within retrieved images to facilitate visualization or a photo montage task similar to the one present in [1], we describe how our bag-of-regions approach achieves satisfactory object localization.

The retrieved region map intuitively suggests the possible location of object even in case of imperfect initial segmentation/retrieval where the region of each object might co-occur with regions of other objects within the image. The object localization task can thus be simplified as detecting the matched region in multiple regions in the retrieved region map.

To achieve this, we adopt a voting scheme based on the repeatability of GF-HOG between sketches and regions, as well

as a consistency measure of bounding boxes aspect ratio. The region with highest voting score S_{vote} is returned as the retrieved object region. Specifically, given the query sketch R_s with a compact bounding box B_s (aspect ratio θ_{B_s}), and region R_k in the retrieved region map with a compact bounding box B_k (aspect ratio θ_{B_k}), the voting score for R_k to be the sketched object is characterized as:

$$S_{vote}(R_k|R_s) = \Phi(\delta_{R_k}, \delta_{R_s}) \cdot \Psi(\theta_{B_k}, \theta_{B_s}) \quad (3)$$

where $\Phi(\delta_{R_k}, \delta_{R_s})$ computes normalized similarity between GF-HOG descriptors in R_s and contour of R_k , and $\Psi(\theta_{B_k}, \theta_{B_s})$ penalizes aspect ratio inconsistency between R_s and R_k .

Specifically, we sample the contours and create putative correspondences $\mathcal{P}_c : \mathcal{P}_m^s \mapsto \mathcal{P}_n^k$ between GF-HOG in R_s and R_k via nearest-neighbor assignment using L^2 norm. Since we have determined the contour of R_k , the sampling and matching errors caused by the clutter from outside the region are significantly reduced compared to [1]. The matching score $\Phi(\delta_{R_k}, \delta_{R_s})$ is computed as:

$$\Phi(\delta_{R_k}, \delta_{R_s}) = e^{-\frac{1}{N} \sum_{\{p_i, p_j\} \in \mathcal{P}_c} \|\delta_{R_s}(p_i) - \delta_{R_k}(p_j)\|^2} \quad (4)$$

where N is the cardinality of the correspondence set.

The region aspect ratio inconsistency penalty $\Psi(\theta_{B_k}, \theta_{B_s})$ is defined as:

$$\Psi(\theta_{B_k}, \theta_{B_s}) = \begin{cases} 1 & \text{if } 0.5 \leq \frac{\theta_{B_s}}{\theta_{B_k}} \leq 2; \\ 0 & \text{else.} \end{cases} \quad (5)$$

3. EXPERIMENT

We evaluate our SBIR system on two datasets: (i) ‘Flickr160’, a dataset published by [1] comprising of 160 creative commons images downloaded from Flickr; five shape categories contains 32 images each. There are 25 free-hand sketches published in the dataset, 5 for each shape class form our query set. (ii) ‘ETHZ Extended Shape Classes’ [17] consists of seven shape categories (apple logo, bottle, giraffe, hat, mug, starfish, swan) in a total of 383 images with around 50 images per category; the 7 sketch models published in the dataset form our queries. The wide range of scales, locations of target objects and the cluttered background make this dataset challenging for global descriptor based SBIR.

The evaluation of our proposed system is performed against the system present in [1] on three leading descriptors GF-HOG [1], SIFT [5] and SSIM [6, 18] for each dataset. We perform queries on configured systems (various combinations of local feature type and codebook size). We compute all descriptors over the edge map (for query sketch) or contour map (for region maps of database images) respectively. Specifically, we compute GF-HOG descriptor with grid size $n = \{5, 10, 15\}$, cell block size $w = 3$, and histogram channels $q = 9$. SIFT is computed on a region of radius 16 pixels. SSIM is computed using a 5×5 correlation window, over a larger 40×40 neighborhood. The SSIM correlation surface is 36 log-polar bins (3 angles, 12 radial intervals).

We compute Average Precision (AP) for each query, averaging over the query set to obtain Mean Average Precision

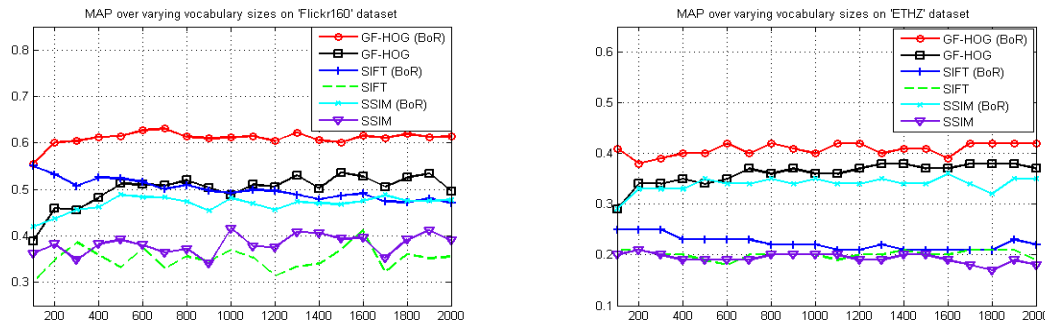


Fig. 4. Performance (MAP) of our system (BoR) against the SBIR system of [1] vs. codebook size (k), comparing three descriptors over the Flickr160 (left) and ETHZ (right) datasets.

(MAP) score. Fig. 4 presents the MAP scores with varying vocabulary (codebook) size k on Flickr160 and ETHZ dataset respectively. For Flickr160, the best performances of using different descriptors on the proposed bag-of-regions based SBIR system are: GF-HOG (63%, $k = 700$); SIFT (55%, $k = 100$); SSIM (49%, $k = 500$). For ETHZ, the best performances are: GF-HOG (42%, $k = 1800$); SIFT (25%, $k = 200$); SSIM (36%, $k = 1600$). Compared to the results reported in [1], we have achieved significant improvement using proposed bag-of-regions (BoR) approach. Specifically, for Flickr160, the best retrieval performance is improved by 9% on GF-HOG, 11% on SIFT and 7% on SSIM; for ETHZ, the best retrieval performance is improved by 4% on GF-HOG, 4% on SIFT and 15% on SSIM. Fig. 1 shows examples of object localization in retrieved images.

4. CONCLUSION

In this paper, we presented a sketch based image retrieval system built on a ‘bag of regions’ (BoR) which encodes the information of shapes at various level of detail in the form of closed contours representing image regions of homogeneous colour. We extended the gradient field HoG and BoVW approach presented in [1] by inferring shape features from region boundaries in the BoR, rather than from Canny edges as in [1] and [2]. The results show that background clutter interferes with shape descriptions only mildly in resulting gradient field. We have demonstrated that the proposed BoR approach yields significant improvements in performance ($\sim 5\%$), for all three descriptors evaluated versus computation over Canny edge maps — and that the GF-HOG descriptor outperforms both SIFT and SSIM in both the Canny/BoVW and BoR frameworks. A further benefit of BoR is we can to localize the sketched object within the matched image almost for free i.e. with only minimal post-processing for clean-up.

5. REFERENCES

- [1] R. Hu, M. Barnard, and J. P. Collomosse, “Gradient field descriptor for sketch based retrieval and localization,” in *ICIP*, 2010, pp. 1025–1028.
- [2] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, “Sketch-based image retrieval: Benchmark and bag-of-features descriptors,” *IEEE TVCG*, vol. 99, 2010.
- [3] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” *IEEE TPAMI*, vol. 5, no. 33, 2010.
- [4] C. Gu, J.J. Lim, P. Arbelaez, and J. Malik, “Recognition using regions,” *CVPR 2009*, pp. 1030–1037, 2009.
- [5] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *IJCV*, vol. 60, pp. 91–110, 2004.
- [6] E. Shechtman and M. Irani, “Matching local self-similarities across images and videos,” in *CVPR*, June 2007.
- [7] J. Ashley, M. Flickner, J. L. Hafner, D. Lee, W. Niblack, and D. Petkovic, “The query by image content (QBIC) system,” in *SIGMOD Conference*, 1995, p. 475.
- [8] J.R. Smith and S.-F. Chang, “Visualeek: a fully automated content-based image query system,” in *ACM Multimedia*, New York, NY, USA, 1996, pp. 87–98, ACM.
- [9] E. Sciascio, M. Mongiello, and M. Mongiello, “Content-based image retrieval over the web using query by sketch and relevance feedback,” in *In Proc. of 4 th Intl. Conf. on Visual Information Systems*, 1999, pp. 123–130.
- [10] C. E. Jacobs, A. Finkelstein, and D. H. Salesin, “Fast multi-resolution image querying,” in *Proc. ACM SIGGRAPH*, Aug. 1995, pp. 277–286.
- [11] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, “A descriptor for large scale image retrieval based on sketched feature lines,” in *SBIM*, 2009, pp. 29–38.
- [12] D. Hoiem, A. A. Efros, and M. Hebert, “Geometric context from a single image,” in *ICCV*, October 2005, vol. 1, pp. 654–661, IEEE.
- [13] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, “Using multiple segmentations to discover objects and their extent in image collections,” *CVPR*, pp. 1605–1614, 2006.
- [14] S. Todorovic and N. Ahuja, “Learning subcategory relevances for category recognition,” *CVPR*, pp. 1–8, 2008.
- [15] P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing,” *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003.
- [16] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*, 2005, pp. 886–893.
- [17] V. Ferrari, “ETHZ extended shape classes,” <http://www.vision.ee.ethz.ch/datasets/>.
- [18] K. Chatfield, J. Philbin, and A. Zisserman, “Efficient retrieval of deformable shape classes using local self-similarities,” *ICCV Workshops*, pp. 264–271, 2009.