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 sketch-based image retrieval system using a bag-of-region representation of images. Regions from the nodes of a hierarchical region tree range in various scales of details. They have appealing properties for object level inference such as the naturally encoded shape and scale information of objects and the specified domains on which to compute features without being affected by clutter from outside the region. The proposed approach builds shape descriptor on the salient shape among the clutters and thus yields significant performance improvements over the previous results on three leading descriptors in Bag-of-Words framework for sketch based image retrieval. Matched region also facilitates the localization of sketched object within the retrieved image.

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

1. INTRODUCTION

Using content keywords to index the explosively growing digital image repository is impractical nowadays, as a complete and accurate set of tags is unfeasible. The proliferation of image data in digital form creates imperative demand for visual content based retrieval techniques. Querying by visual example (QVE) provides an attractive alternative to keyword based approach. Recent successes with bag-of-visual-words (BoW) approaches have led to photo-realistic query systems exhibiting leading performance in benchmark tasks (e.g. PASCAL) and scalability over large image collections.

The BoW approaches have also been adopted and proved to be successful in image retrieval systems using sketch queries in more recent works [1] [2]. However, both systems propose shape descriptors based on low level visual information extracted from photo-realistic images, e.g. strong edge map from Canny operator, without considering middle level visual information. In many cases, the shape information can not be precisely captured from high frequency information due to the cluttered background. While strong edge is a fundamental cue for sketch based image retrieval (SBIR), it can often be too weak of a signal to reliably detect object in texture rich images. The recent advances in contour detection and hierarchical segmentation [3] has shed light on object recognition using regions [4] which are more spatially extended and perceptually meaningful entities than low-level interest point-based features [5]. Regions have appealing properties for object level inference such as the naturally encoded shape and scale information of objects and the specified domains on which to compute features without being affected by clutter from outside the region.

This paper proposes to use the bag-of-regions approach decomposing images as region representations for sketch based image retrieval. Regions are collected from the nodes of a hierarchical region tree generated by [6], and these regions range in scale from super pixels to the whole image. We extend the work present in [1] inheriting the Gradient Field HOG (GF-HOG) descriptor and BoW paradigm, while adopting a hierarchical tree of regions to extract object contours ranging in various scales (sec. 2). In general, contour detectors offer no guarantee of producing object contours suppressing high frequency information (e.g. texture) whilst closed object contours can always be recovered from regions in the form of their boundaries. The proposed bag-of-regions approach yields significant improvements in performance compared to the results reported in [1] based on three leading descriptors and BoW framework: GF-HOG, SIFT, and Self-Similarity Descriptor (SSIM) (sec. 3). Moreover, matched region also facilitates the localization of sketched object within the retrieved image (sec. 2.4).

1.1. Related Work

Early sketch based image retrieval systems are typically driven by queries comprising blobs of color or predefined texture [7] [8], which are later augmented with shape descriptors [9] and spectral descriptors such as wavelets [10]. Eitz *et al.* [11] introduce grid approaches divide the image into regular grids and locate photos using sketched depiction of object shape. Descriptors from each cell are 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 is proposed by Hu *et al.* [1]. This descriptor is able to mitigate the lack of relative spatial information within BoW by capturing structure from surrounding regions. Eitz *et al.* [2] improve the standard HOG descriptor by storing only the most dominant sketched feature lines in the HOG and adopt a similar BoW approach.

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

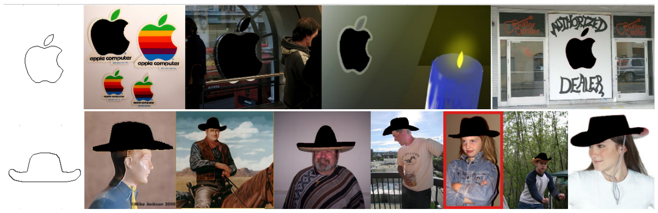

Fig. 1. Sample query sketches and some of the top returned images. The localized objects are marked as black in the image.

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] develop 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].

2. SYSTEM OVERVIEW

The pipeline of the system is as follows: First, each image is represented by a bag of regions derived from a region tree as shown in Figure 2. Next, we extract the contours of the region maps containing various levels of detail, and capture local structure in the contour map using GF-HOG descriptor. After that, the shape descriptors are clustered to form a BoW codebook via k-means and the resulting histogram is constructed for matching. In the final step, 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 present in [6] as shown in Fig. 2 . We discard the tree structure and take all the nodes of that tree, except the root which is the entire image plane. Intuitively, this bag of regions stores visual coherent regions at various details. Specifically the contours and corresponding segmentation are obtained by thresholding the Ultrametric Contour Map (UCM) at level 0.4.

At this stage, we are not assuming that each object can be correctly segmented in any hierarchy, given the fact that the region of each object might frequently co-occur with regions of other objects within the image. Although not all objects might have exclusive region map in the bag of regions, the best corresponding region map of the object in the view of least co-occurring regions of other objects encodes the shape and scale information of this object without being significantly affected by clutter from outside the object.

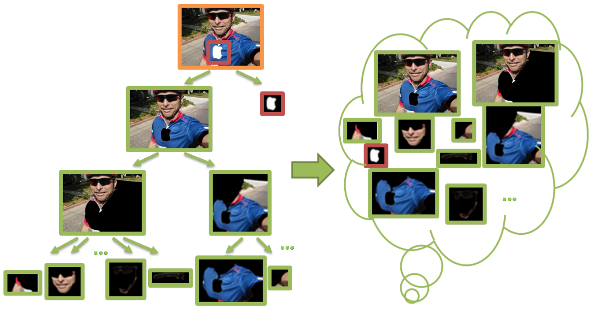

Fig. 2. The "bag of regions" representation of a "biker" example. Regions are collected from all nodes of a region tree generated by [6], ranging in various levels of details.

2.2. Descriptor

Constructing a BoW codebook 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*. HOG descriptor built on the dense *gradient field*, i.e. GF-HOG is reported to outperform SSIM and SIFT. We adopt GF-HOG in our system but evaluation results on SSIM and SIFT descriptors are also given 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:

$$\underset{\Theta}{\text{argmin}} \int\int_{\Omega} (\triangledown\Theta - \mathbf{v})^2 \quad s.t. \;\; \Theta|_{\delta\Omega} = \theta|_{\delta\Omega}. \tag{1}$$

i.e. $\triangle\Theta = 0$ 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 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 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.

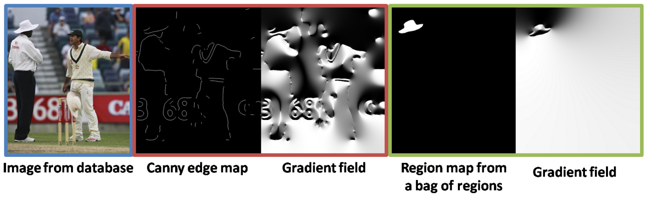

Fig. 3. *Gradient field* of a region map from the bag of regions (green window) and the Canny edge map (red window). The *Gradient field* of the region map captures complete information of the salient shape among the clutters, i.e. the hat in this example.

2.3. BoW Framework for Sketch Based Retrieval

We learn a BoW codebook by clustering GF-HOG descriptors from the bag of regions of all images via k-means. We construct a frequency histogram H^I_m 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 $d(H^I_m, H^s)$.

$$\begin{aligned} d(H^S, H^I_m) &= \sum_{i=1}^{k}\sum_{j=1}^{k}(\omega_{ij}\min(H^S(i), H^I_m(j))), \\ \omega_{ij} &= 1 - |\mathcal{H}^S(i) - \mathcal{H}^I_m(j)|. \end{aligned} \tag{2}$$

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 segmentation 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 it, we adopt a voting scheme based on the repeatability of GF-HOG between sketches and regions, as well as a consistence measure of bounding boxes aspect ratio. The region with highest voting score 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:

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

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

Specifically, we sample the contours and create putative correspondences $\mathcal{P}_c : \mathcal{P}^s_m \mapsto \mathcal{P}^k_n$ between GF-HOG in R_s and R_k via nearest-neighbor assignment using L^2 norm. Since we have the determinate 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:

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

where N is the cardinality of the correspondence set.

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

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

3. EXPERIMENT

We evaluate our SBIM 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' [16] 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 very challenging for the common global descriptor based SBIR systems.

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 [17, 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 partitioned into 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 (MAP) score. Fig. 4 and Fig. 5 present 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

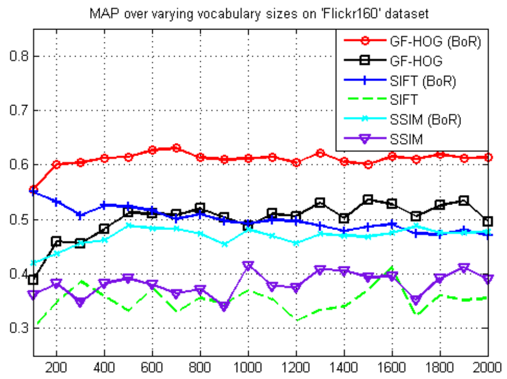

Fig. 4. Performance (MAP) of our system (BoR) against the SBIR system present in [1] vs. codebook size, comparing three descriptors over Flickr160.

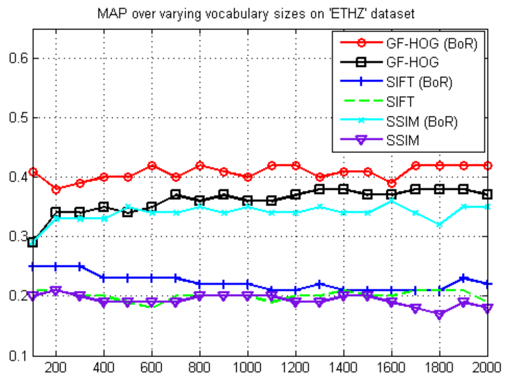

Fig. 5. Performance (MAP) of our system (BoR) against the SBIR system present in [1] vs. codebook size, comparing three descriptors over ETHZ sets.

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 which encodes the complete information of salient shapes at various level of details in the form of enclosed contours of regions presenting a coherent visual appearance. We extended the gradient field and BoW approach present in [1], by inferring shape features from object contours in a bag-of-regions paradigm. The results show that background clutter interferes with shape descriptions only mildly in resulting gradient field. We have demonstrated that the proposed bag-of-regions approach yields significant improvements in performance, outperforming the results reported in [1] based on three leading descriptors and BoW framework: GF-HOG, SIFT and SSIM. Our system is also able to localize the sketched object within the retrieved image facilitated by the matched region.

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.C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," 2009, pp. 2294–2301.

[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] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE TPAMI*, vol. 99, 2010.

[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, "Visualseek: 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] V. Ferrari, "ETHZ extended shape classes," `http://www.vision.ee.ethz.ch/datasets/`.

[17] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *CVPR*, June 2007.

[18] K. Chatfield, J. Philbin, and A. Zisserman, "Efficient retrieval of deformable shape classes using local self-similarities," *ICCV Workshops*, pp. 264–271, 2009.