GRADIENT FIELD DESCRIPTOR FOR SKETCH BASED RETRIEVAL AND LOCALIZATION

Rui Hu, Mark Barnard and John Collomosse

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

ABSTRACT

We present an image retrieval system driven by free-hand sketched queries depicting shape. We introduce Gradient Field HoG (GF-HOG) as a depiction invariant image descriptor, encapsulating local spatial structure in the sketch and facilitating efficient codebook based retrieval. We show improved retrieval accuracy over 3 leading descriptors (Self Similarity, SIFT, HoG) across two datasets (Flickr160, ETHZ extended objects), and explain how GF-HOG can be combined with RANSAC to localize sketched objects within relevant images. We also demonstrate a prototype sketch driven photo montage application based on our system.

Index Terms— Sketch based Image Retrieval, Bag-of-visual-words, HoG, RANSAC.

1. INTRODUCTION

Digital image libraries are commonly indexed using content keywords, yet many creative applications call for the retrieval of imagery based on visual appearance. The task of querying databases by visual example (QVE) has received considerable attention in recent years, with *bag-of-visual-words* approaches to QVE exhibiting leading performance in benchmark tasks (e.g. PASCAL) and scalability over many thousands of images or video frames. However such systems require a photo-realistic query which, in many use cases, may be unavailable to the user. This paper presents a codebook based QVE system capable of both retrieving images using *free-hand sketched queries* depicting object shape, and *localizing the position* of the sketched object within those images.

Bag-of-visual-words (BoW) techniques create a codebook of visual words from discriminative features (*descriptors*) local to points within database images. Both the images and query are described using frequency histograms of visual words present. To perform retrieval, the histogram of the query is compared to those of the database. Descriptors must therefore exhibit high repeatability across the query image and relevant images in the database.

Although the scalability of BoW is attractive, adapting the approach to sketch based retrieval is challenging. Sketched objects do not share the rich photometric properties of their image counter-parts, and object depictions often differ in scale or location, and may be subject to non-linear shape deformations. Moreover, sketches are visual structures defined by the spatial inter-relationships of a sparse set of strokes. The BoW representation (a global histogram of locally sampled descriptors) lacks information on the spatial distribution of descriptors. This is appropriate for photos — where descriptors capture *rich* information and spatial relationships are *less* important — but not for sketches, where the converse is true.

This paper introduces the Gradient Field HoG (GF-HOG) descriptor; an adaptation of HoG that mitigates the lack of relative spatial information within BoW by capturing structure from surrounding regions. We are inspired by work on image completion (in-painting) capable of propagating image structure into voids, and use a similar "Poisson filling" approach to improve the richness of information in the gradient field prior to sampling with the HoG descriptor. This simple technique yields significant improvements in performance when matching sketches to photos, compared to three leading descriptors: Self-Similarity Descriptor (SSIM); SIFT; and HoG (sec. 2.1). Furthermore we show how the descriptor can be applied to localize sketched objects within the retrieved images (sec. 3), and demonstrate this functionality through a sketch driven photo montage application (sec. 3.1).

1.1. Related Work

Sketch based Image Retrieval (SBIR) arguably began to gain momentum in the mid-nineties with blob based retrieval systems, such as QBIC [1] and VisualSeek [2]. Queries comprise blobs of color or texture; spatial distribution is modelled using wavelets [3], or dividing the image into a regular grid and matching cells using local color or texture descriptors [1]. Grid approaches have also been used to locate photos using sketched depiction of object shape (via EHD [4] or structure tensor [5]). Descriptors from each cell are concatenated to form a global image feature. However this offers limited invariance to changes in position, scale or orientation.

Contour descriptors have been used to match sketched shapes to images. Edge segments are tokenized into a string representation, encoding length, curvature, and relative spatial relationships [6]. Edge orientation [7, 4] and angular partitioning [8] have also used to describe contours. Model fitting approaches such as [9] deform the sketch to fit to edges of objects in the images, measuring similarity via the deformation energy spent. Although these more expensive approaches offer improved tolerance to depictive inaccuracy, they do not scale well over large databases.

Shechtman and Irani proposed Self-similarity (SSIM) as an image descriptor invariant to depictive style [10]. Recently, Chatfield *et al.* [11] reported experiments using SSIM in a BoW framework to retrieve images using photo-realistic queries of shapes. In sec. 4 we evaluate SSIM, alongside SIFT and HoG descriptors for the purpose of sketch based shape retrieval and show our adapted GF-HOG descriptor outperforms these on both our own Flickr160 dataset[1] and the ETHZ extended object dataset (a subset of which is used in [11]).

2. GRADIENT FIELD DESCRIPTOR

Our system accepts monochrome free-hand sketched queries depicting a shape, and returns images that contain similar shapes. This requires a matching process robust to depictive inaccuracy (e.g. in location, scale, or shape deformation) and photometric variation. Our approach is to transform database images into Canny edge maps, and capture local structure in the map using a novel descriptor (subsec. 2.1). We recommend setting an appropriate scale and hysteresis threshold for the Canny operator by searching the parameter space for a binary edge map in which a small, fixed percent of pixels are classified edge. This simple heuristic extracts dominant edges and discourages response at the scale of finer texture.

2.1. Gradient field

Shape and structure (both in the sketches and Canny maps) are encoded in the relative location and spatial orientation of edges. Constructing a BoW codebook using local descriptors such as SSIM, SIFT, HoG results in poor retrieval performance, as we later show (sec. 4.1). 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 solution is to represent image structure using a dense *gradient field*, interpolated from the sparse set of edge pixels. Given an edge map $M(x,y) = \{0,1\}$, we compute a sparse field from the gradient of edge pixels $G[x,y] \mapsto \operatorname{atan}\left(\frac{\delta M}{\delta x} / \frac{\delta M}{\delta y}\right), \forall_{x,y} M(x,y) = 1$. We seek a dense field $\mathcal{G}_\Omega$ over image coordinates $\Omega \in \Re^2$ constrained such that $\mathcal{G}(p) = G(p), \forall_{p \in \Omega} M(p) = 1$, and minimizes the energy:

$$\underset{\mathcal{G}}{\operatorname{argmin}} \int\int_\Omega \left(\bigtriangledown \mathcal{G} - G\right)^2 \quad s.t. \;\; \mathcal{G}|_{\delta\Omega} = G|_{\delta\Omega}. \qquad (1)$$

i.e. $\triangle \mathcal{G} = 0$ over Ω s.t. $\mathcal{G}|_{\delta\Omega} = G|_{\delta\Omega}$ for which a discrete solution was presented in [12] solving Poisson's equation with Dirichlet boundary conditions. This can be approximated by forming a set of linear equations for non-edge pixels, that are fed into a sparse linear solver to obtain the complete field. Common applications such as image completion ("Poisson in-filling" [12]) approximate $\triangle \mathcal{G} = 0$ using a 3×3 Laplacian window: $4\mathcal{G}(x,y) = \mathcal{G}(x-1,y) + \mathcal{G}(x+1,y) + \mathcal{G}(x,y-1) + \mathcal{G}(x,y+1)$. However we obtained better results in our retrieval application when approximating $\triangle \mathcal{G}$ using a 5×5 window discretely sampling the Laplacian of Gaussian operator (leading to a smoother field):

$$\triangle \mathcal{G}(x,y) = -\frac{1}{\pi\sigma^4}\left[1 - \frac{x^2+y^2}{2\sigma^2}\right] e^{-\frac{x^2+y^2}{2\sigma^2}}. \qquad (2)$$

Images and sketches are padded with an empty border of 15% pixel area. For typical images of $\sim 200 \times 100$ the linear system is solvable in ~ 0.7s using TAUCS [13].

[1] Available at: **http://personal.ee.surrey.ac.uk/Personal/R.Hu/ICIP**

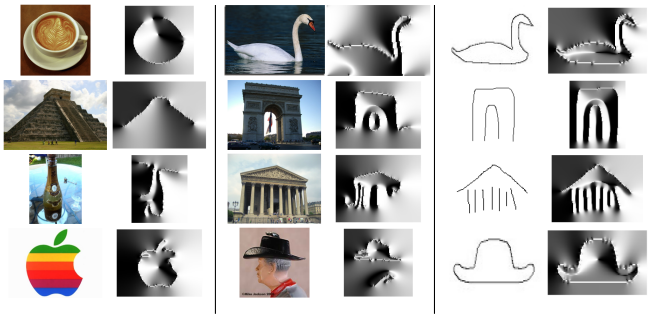

Fig. 1. Sample images and query sketches. Corresponding visualizations of field $\mathcal{G}$ following processing of subsection 2.1.

2.2. Multi-scale Histogram of Gradient

The Histogram of Gradient Orientation (HoG) descriptor [14] is widely applied in object classification, and human (e.g. pedestrian) detection. The descriptor is computed within a window centered upon a point (either key-point of interest or densely sampled). The window is divided into a regular $n \times n$ grid, and a frequency histogram is constructed within each grid cell according to the edge orientation of pixels within. To enable histogram construction, the range of edge orientations is quantized into q bins. The histogram counts are concatenated to form a q-D vector for each cell, which are again concatenated to form an qn^2-D vector for the window. In many implementations, several windows are sampled in a non-overlapping $w \times w$ grid local to the key-point and again concatenated to output the final descriptor.

Our system computes a HoG descriptor at $\mathcal{G}(x,y)$ for all points where $M(x,y) = 1$. (i.e. pixels comprising sketched strokes, or in the case of database images, Canny edges). To cope with problems of scale change we detect HoG features with $n = \{5, 10, 15\}$ and fix $w = 3$, $q = 9$, yielding several hundred Gradient Field HoG (GF-HOG) descriptors for a typical image. Although a multi-scale approach is also adopted by Pyramid HoG (PHOG) [15] (photo queries), descriptors at each scale are concatenated to form a feature whereas we add GF-HOG at each scale (n) to the image descriptor set.

2.3. Sketch based retrieval using BoW

GF-HOG from all images are clustered to form a BoW codebook via k-means (sec. 4 presents results for varying k). A frequency histogram H^I is constructed for each image. At

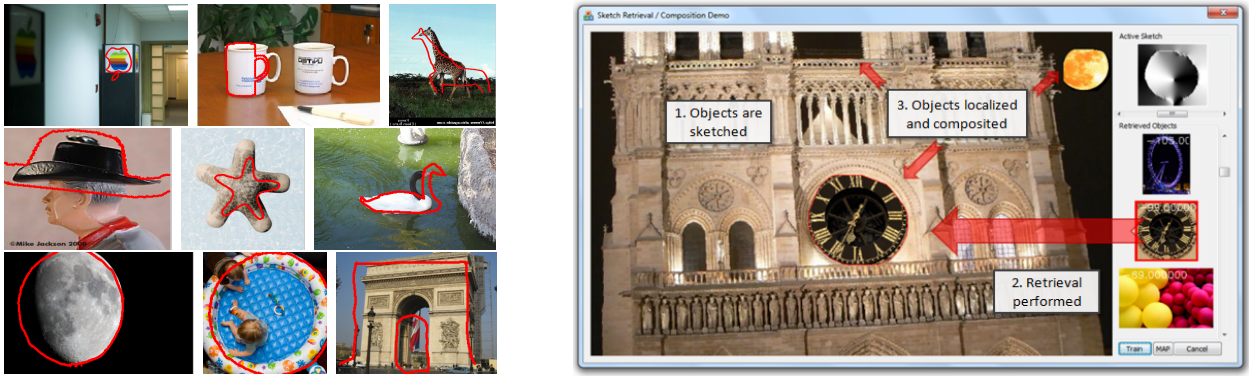

Fig. 2. Left: Localizing the query sketch within retrieved images. Right: Photo montage driven by our BoW retrieval system.

query time, a frequency histogram H^s is constructed from the query sketch by quantizing GF-HOG from the sketch using the same codebook. Images are ranked according to histogram similarity $d(H^I, H^s)$. A common choice for $d(.)$ is an L^2 norm; efficiently computable over an n member dataset via kd-tree in $O(\log n)$. We have found histogram intersection to outperform L^2 in average precision, though this improvement incurs increased complexity of $O(n)$.

$$\begin{aligned} d(H^S, H^I) &= \sum_{i=1}^{k}\sum_{j=1}^{k} \min(\omega_{ij} H^S(i) H^I(j)), \\ \omega_{ij} &= 1 - |\mathcal{H}^S(i) - \mathcal{H}^I(j)|. \end{aligned} \quad (3)$$

where $H(i)$ indicates the i^{th} bin of the histogram, $\mathcal{H}(i)$ the normalized visual word corresponding to the i^{th} bin.

3. OBJECT LOCALIZATION

For visualization of results, and for our photo montage prototype (subsec. 3.1), it is desirable to compute the position of the sketched shape within retrieved images. Given typical perturbations of object shape within a sketch query, any localization is expected to be approximate. Here, we demonstrate the repeatability of GF-HOG between sketches and photos by applying RANSAC to fit the sketched shape to the image via a rigid transformation. We simplify by modelling the transformation as a linear conformal affine transform (LCAT); i.e. a uniform scale, a rotation and a translation. Two points of correspondence are required to define an LCAT.

Given a sketch and a retrieved image, we first create putative correspondences between GF-HOG in the sketch and the image via nearest-neighbor assignment using L^2 norm. If descriptor A in a sketch is assigned to B in the image, then B must also be nearest to A for a valid match [16]. Given putative correspondences $\mathcal{P}_m^S \mapsto \mathcal{P}_n^I = \{p_{i=1\ldots m}, p_{s=1\ldots n}\}$, our iterative search runs as follows. We randomly sample two pairs of correspondences, deducing the LCAT (T). We then compute the transfer error $E(T)$ using all correspondences:

$$E(T) = \sum_{\{p_s, p_i\} \in \mathcal{P}^S \mapsto \mathcal{P}^I} |p_s - T p_i|^2 + |p_i - T^{-1} p_s|^2 \quad (4)$$

We iterate for up to 10,000 trials seeking to minimize $E(T)$. Fig. 2 (left) visualizes typical localized results.

3.1. Sketch based photo montage

We have applied our retrieval and localization algorithms to develop a prototype system for photo montage using sketched queries (Fig. 2, right). The system is similar in spirit to Chen *et al.*'s Sketch2Photo [17], except that we use sketched shape to retrieve our images rather than running a keyword search. Users sketch objects and select photos to insert into the composition from ranked results on the right. Upon selecting an image, the position of the sketch shape is localized and the region of interest cropped and composited into the sketch. Unlike Eitz *et al.*'s PhotoSketch [5] our GF-HOG enables matching invariant to scale and position of sketched shapes.

4. EXPERIMENT

We evaluate our system over two datasets: (i) 'Flickr160' a dataset of 160 creative commons images downloaded from Flickr, comprising five shape categories contains 32 images each. Our query set comprises 25 free-hand sketches; 5 for each shape class (e.g. Fig. 1, right). (ii) 'ETHZ Extended Shape Classes' [18] is a standard shape dataset with 383 images, comprising seven shape categories (apple logo, bottle, giraffe, hat, mug, starfish, swan) containing around 50 images per category, the 7 sketch models published in the dataset form our queries.

4.1. Comparative Evaluation of GF-HOG

For each dataset we perform BoW retrieval using the proposed GF-HOG descriptor. We compare retrieval performance with an otherwise identical system incorporating alternative descriptors: the Self-Similarity (SSIM) [10, 11], SIFT [16], and HoG [14] descriptor. In all cases descriptors are computed over the edge map (for database images) and sketched image as appropriate.

Here, SIFT is computed on a region of radius 16 pixels. SSIM is computed using a 5×5 correlation window, over a

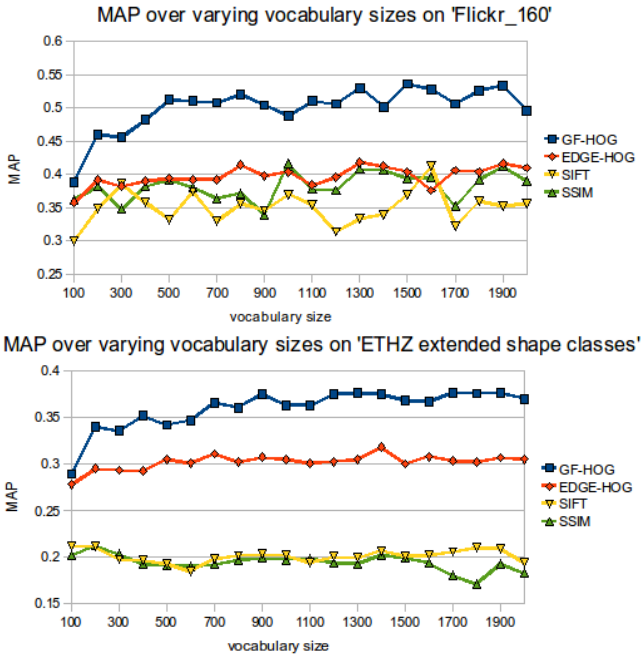

Fig. 3. Performance (MAP) of our system vs. codebook size, comparing four descriptors over Flickr160 and ETHZ sets.

larger 40×40 neighborhood. The SSIM correlation surface is partitioned into 36 log-polar bins (3 angles, 12 radial intervals). HoG is computed with identical parameters to our GF-HOG descriptor (subsec. 2.1), i.e. we compute over multiple window scales for fairness as this outperforms a single scale classic HoG [14] in all our test cases. By computing HoG over the edge image, and GF-HOG over gradient field, we directly show the benefit of our gradient field over multi-scale classic HoG (referred to here as EDGE-HOG).

Average Precision (AP) is computed for each query, and averaged over the query set to produce Mean Average Precision (MAP) score. Fig. 4 presents these over a range of vocabulary (codebook) size k. For Flickr160, the best performances are: GF-HOG $(54\%, k = 1500)$; EDGE-HOG $(42\%, k = 1300)$; SIFT $(41\%, k = 1600)$; SSIM $(42\%, k = 1000)$. For ETHZ, the best performances are: GF-HOG $(38\%, k = 1300)$; EDGE-HOG $(32\%, k = 1400)$; SIFT $(21\%, k = 100)$; SSIM $(21\%, k = 200)$. The trends of Fig. 3 show significant improvement using GF-HOG over contemporary descriptors of $\sim 10\%$ for Flickr160 and $\sim 5\%$ for ETHZ. Examples of localization are given in Fig. 2 (left).

5. CONCLUSION

We have shown that the proposed GF-HOG descriptor can be effectively incorporated into a BoW system for sketch based image retrieval. To the best of our knowledge BoW has not been previously used to retrieving images using free-hand sketched shapes. Furthermore, our descriptor out-performs SSIM, SIFT and HoG descriptors for this task. We have also demonstrated a prototype application for our retrieval technique (sec. 3.1). The success of the descriptor is dependent on correct selection of scale during edge extraction, and use of image salience measures may benefit this process. The system could be enhanced by exploring colored sketches, or incorporate more flexible models for object localization.

6. REFERENCES

[1] 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.

[2] 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.

[3] C. E. Jacobs, A. Finkelstein, and D. H. Salesin, "Fast multiresolution image querying," in *Proc. ACM SIGGRAPH*, Aug. 1995, pp. 277–286.

[4] J.-L. Shih and L.-H. Chen, "A new system for trademark segmentation and retrieval," *Image and Vision Computing*, vol. 19, pp. 1011–1018, 2001.

[5] 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.

[6] Y. Chans, Z. Lei, D. Lopresti, and S. Y. Kung, "A feature-based approach for image retrieval by sketch," in *SPIE Storage and retrieval for image and video databases*, 1997.

[7] A. K. Jain and A. Vailaya, "Image retrieval using color and shape," *Pattern Recognition*, vol. 29, pp. 1233–1244, 1996.

[8] A. Chalechale, G. Naghdy, and A. Mertins, "Sketch-based image matching using angular partitioning," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 35, pp. 28–41, 2005.

[9] A. Del Bimbo and P. Pala, "Visual image retrieval by elastic matching of user sketches," *IEEE PAMI*, vol. 19, no. 2, pp. 121–132, 1997.

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

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

[12] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003.

[13] S. Toldeo, D. Chen, and V. Rotkin, "TAUCS: A library of sparse linear solvers," http://www.tau.ac.il/ stoledo/taucs/.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886–893.

[15] A Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *CIVR*, New York, NY, USA, 2007, pp. 401–408, ACM.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[17] T. Chen, M.-M. Cheng, P. Tan, S. Ariel, and S.-M. Hu, "Sketch2Photo: internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 1–10, 2009.

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