# Object Search via Random Context

Wei Liu
Computer School of Huazhong University of Science and Technology
Computer School of Huazhong Normal University
Wuhan ,China
Email:liuwei@mail.ccnu.edu.cn

Yunxing Ruan
Computer School of Huazhong Normal University
Wuhan ,China
Email: ruan_yx@mail.ccnu.edu.cn

Xia Cai
Computer School of Huazhong Normal University
Wuhan ,China
Email: caixia@mail.ccnu.edu.cn

**Abstract—Accurately object searching plays an important role in computer vision. Retrieving and locating target objects in images are object searching's two sub-tasks. Aiming to promote the precision and recall of object searching, selecting appropriate image representation methods is the core issues. The representation method needs to provide enough discriminative features. Our approach adopts locality sensitive hashing method to extract enough sift features. The extracted features contain inliers and outliers. In order to distinguish them, random context confidence scores of features are computed. Our algorithm offers 3 benefits:1) A novel partition method is adopted to divide images. It is easy to be parallelized during computing contexts.2) A novel random points selecting method is adopted to avoids ill-defined boundary for target objects; 3) Multiple target objects in one image can be located by clustering all the features of each image with their coordinates. The experiment on a challenging Belgalogo dataset highlights the performance of our approach.**

***Keywords- Object Search; Visual Word; Random Context; Multiple Objects Location***

## I. INTRODUCTION

Despite the great development in the field of image understanding, object searching is still a challenging problem in computer vision, due to occlusion, illumination, non-rigid distortion and other reasons. The most common way of object searching is to select appropriate features to represent images, and then match these features against the query objects' ones. Similar to natural language retrieval, the discrimination of single feature is not good enough. Many works adopt context or weighted information to promote the single feature's performance. However, how to choose and quantize the contexts are still problems in information retrieval.

Currently, there are three methods for defining contexts for interest points: the first category of methods[1, 2] uses image segmentation or region detection results as contexts of features. The performance is highly depending on the accuracy of the image segmentation results. The second category of method binds each interest point with its recurrent neighbors as its context, called phrase[3]. Fixed phrases promote features' discrimination effect greatly. But it depends on the prior knowledge of the size of phrases. The third category selects a fixed size window[4], which contains interest points defined as its contexts. But the fixed size of the window is also hard to select.

In order to avoid image segmentation and parameter selection, we propose a novel method by random dividing each image into several partitions, as illustrate in Fig. 1. Every interest point must be located in one partition of the image. The histogram of the features in each partition is computed, which is defined as the interest point's context in one partition. After several times partitions, we compute the exception of histograms of all patches as each interest point's context confidence score. Thus, the algorithm can be compute paralleled, and do not need to fix the size of the neighbor window.

The boundary points of target objects share the same context with their neighbors in most of time. And this causes ill-defined boundary. Technically, we put the split points in the region of interest points to avoid the problem.

In real world, one image contains 0, 1 or more than 1 target objects. Finding all the bounding boxes in an image is another challenging problem. Intuitively, the interest points of one target object are always gathered together. They can be clustered into serval groups according to their coordination. Once the number of interest points in one group is enough, a target object can be determined by this group of interest points. So, by clustering interest points, the recall of the object searching can be promoted greatly.

The remainder of this paper is organized as follows. In Section2, the related work about image representation and object location is discussed. Section3 presents the details of our method, including interest points' extraction by locality sensitive hashing method, random partition of images, and object location by RANSAC. The experiments and data analysis are in section 4. We conclude this paper in Section 5.

## II. RELATED WORK

Image representation is the basis of object searching. The interest points' representation methods determine their context representation methods. Generally, the interest points' representation methods can be classified into two categories. One is the low-level feature representation, such as shape, texture or color. Accordingly, the contexts are represented by shape context[5], texture context[6], and color context[7], etc. The second category is middle-level features, such as visual bag-of-word method[8, 9]. The middle-level features always integrate several low-

level features and quantize them. Compared with the low-level features, the middle-level features are more abstract and briefer in image classification, like words and roots working in the natural language understanding. As discussed in [10], the middle-level feature causes ambiguity. Most algorithms focus on promoting the discrimination of visual words. For example, kernel functions are trained in supervised learning, and the histograms of visual phrases worked as contexts are computed in unsupervised learning.

Object searching also includes target objects' location. Its task is to reveal the object's position and size with rectangular, ellipse or irregular geometric figures etc. Given interest points' coordinate, including inliers and outliers, RANSAC algorithm[11] can be used to locating the object with rectangle by computing the optimal affine transformation model. But when the target object is occluded, the calculated rectangle for the target will be larger than its actual size. Efficient Sub-image Retrieval algorithm[12]computes the max confidence score window by branch and bound search method in sub-linear time. However it is unable to locate several targets in an image. Voting method[13] is also used in locating target objects by computing every pixel's confidence score. The threshold is set to determine whether the pixel is part of target object or not. This method is robust in occluded and distorted target location except that the parameter is hard to select.

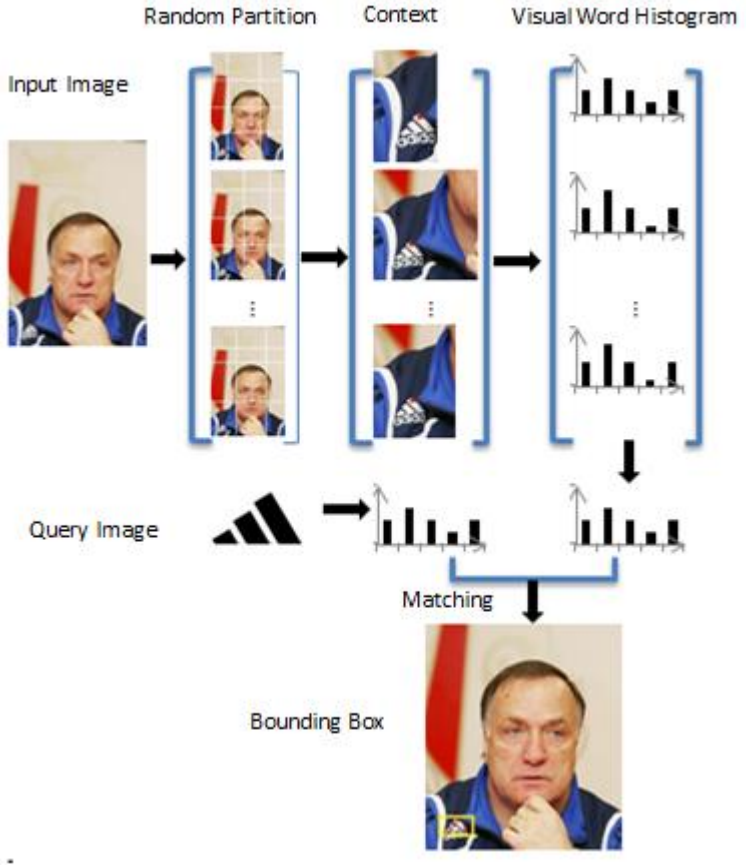


Figure 1. Illustrate the random partition for every interest point. An interest point is shown in red in the input image. After dividing, the contexts contained the interest point is extracted, and the histogram can be computed. Matched against the query image's histogram, confidence score of the interest point is computed. Given the confidence scores, the interest points are divided into outliers and inliers. By RANSAC algorithm, the bounding boxes of the target objects can be draw.

## III. RANDOM CONTEXT

### A. *Interest Points Extraction*

Since SIFT feature is scale and rotate invariant, a lot of image feature matching algorithms use SIFT or SIFT like features to represent images. But experiments show that simply matching SIFT features leads to poor result.

The k nearest neighbors for query image's SIFT points are defined as interest points. The interest points shows a better performance [14], especially in a large dataset. Obviously, the K nearest interest points include inliers and outliers. The outliers must be eliminated before object matching.

### B. *Random Partitions*

We adopt Random Context to eliminate the outliers from interest points. As Fig. 1 illustrates, each image is divided into M*N partitions. Every partition is noted as $P_i, i \in \{1, .., M * N\}$ . Every interest point locates in one partition. Then the partition is defined as the interest point's context in this division. Given each interest point, the histogram of all the SIFT feature in the partition $h_{P_i}$ , is computed to express its context. And then it is matched against the target object's feature histogram, $h_{query}$.

Repeating the division for K times, every interest point in the image can be assigned K confidence scores. The similarity score $S(P_i)$ is defined as expectation of the

confidence score of the partition $P_i$ against the query object:

$$S(P_i) = E\left(score(P_i, query)\right) = \sum_{k=1}^{K}(p_i * score_k(P_i, query)) \approx \frac{1}{K}\sum_{k=1}^{K} score_k(P_i, query) = \frac{1}{K}\sum_{k=1}^{K} distance_k\left(h_{P_i}, h_{query}\right) \quad (1)$$

The higher confidence score is, the more similar between the partition and the target object. We consider two interest points $i \in P_i$ and $j \in P_j$. The point i is located inside the ground truth region G, $i \in G$; and the point j is located outside G, $j \notin G$. After K times, the confidence score of point i must be greater than that of point j:

$$\lim_{K\to\infty} P\left(S^K(P_i) > S^K(P_j)\right) = 1 \quad (2)$$

We denote $F(P)$ as the features set. All the SIFT features in $F(P)$ are located inside the partition P. Suppose the score of partition P is linearly proportional to the number of the element in the intersection set, $|F(P) \cap F(G)|$. The more features belonged to the intersection of P and G, the more similar P and G is. That is:

$$S(P) \propto |F(P) \cap F(G)| \quad (3)$$

The number of $F(P) \cap F(G)$ is also linearly proportional to the area of $P \cap G$:

$$|F(P) \cap F(G)| \propto Area(P \cap G) \quad (4)$$

[12] has proved that $E(Area(P_i \cap G)) \geq E(Area(P_j \cap G))$. So, the interest point inside the target object has the higher score than the other points as shown in formula 2.

### C. object location

After assigning each interest point's confidence score, the interest points with higher score are left to locate the bounding box for target objects, and the interest points with lower scores are eliminated, which are defined as outliers. In real world, the image contains several target objects (n=0, 1..). Our task is to locate all the query objects in images. Intuitively, the interest points belonged to one object are gathered together. By clustering them, all the interest points are divided into several groups, according to their coordinates. Thus, each group corresponds to an affine transformation matrix using RANSAC algorithm. In the end, the overlapped boxes are merged for they contain the same target object in the image.

## IV. EXPERIMENT

The test dataset Belgalogo[15] is composed of 10,000 images covering all aspects of life and current affairs, all images have been manually annotated for 26 logos. The query images in Qset1 are composed of 55 image patches from the dataset images. From these images, we extract a total of 24,172,440 sift points and their coordinates.

### A. Results Evaluation

To evaluate the performance, we calculate recall and precision score for logos searching in the Belgalogo dataset. The recall of the result is the percentage of correct detections against the total number of ground truth images. Precision is the fraction of retrieved images that are relevant. The correct (or relevant) detections are defined as those boxes whose differences are less than 20 pixels compared with the ground truths' coordinates. In experiment, the number of nearest neighbor k is 50, the number of the split points is 5. The size of visual word dictionary is 200,000. The distance function is Euclidean distance, the threshold is 35. In order to locate all the bounding boxes in each image, the interest points in one image are divided into n groups. During the clustering, the number of groups varies from 1 to n, keeping each group contains more than 3 interest points. the reason is that the RANSAC algorithm need 3 points at least to compute the affine model. The results are shown in Table 1.

The result shows that the recall score of "President" is less than that of K-nn algorithm. And the other recall scores are better than those of K-nn. the reason is that many images contain more than one logos, and most "President" logos only appear in one image. After clustering the interest points, more target logos can be found in the images. So, the recall score is higher. And after filtering by the contexts confidence score, the precision scores are almost all better than those of K-nn and ESR.

TABLE I. THE RECALL OF OBJECT SEARCHING WITH K=50 AND N=6

| Logo Name | K-nn | ESR | RC | clustered _RC |
|---|---|---|---|---|
| Base | 0.154 | 0.025 | 0.160 | 0.160 |
| Dexia | 0.238 | 0.024 | 0.328 | 0.332 |
| Ferrari | 0.730 | 0.026 | 0.676 | 0.811 |
| Kia | 0.340 | 0.177 | 0.340 | 0.645 |
| President | 0.714 | 0.643 | 0.643 | 0.643 |

TABLE II. THE PRECISION OF OBJECT SEARCHING WITH K=50 AND N=6

| Logo Name | K-nn | ESR | RC | clustered _RC |
|---|---|---|---|---|
| Base | 0.455 | 0.0714 | 0.500 | 0.500 |
| Dexia | 0.463 | 0.429 | 0.448 | 0.481 |
| Ferrari | 0.771 | 0.02 | 0.325 | 0.316 |
| Kia | 0.495 | 0.1656 | 0.366 | 0.615 |
| President | 0.769 | 0.09 | 0.9 | 0.9 |

Some object searching results are shown in the Fig. 2, including KIA, President, Base and Dexia etc. The solid green boxes are the location of the target objects. The result shows that several target objects in an image can be located after clustering interest points. The first image in Fig. 2 shows two boxes in locating the logo KIA. The bottom one in the first image locates a occluded KIA logo. Because our approach use the RANSAC algorithm to locate the logo, the box is greater than the real target object. It is the weak point of the algorithm.

Figure 2. Examples of Object Location

### B. *Selecting Splitting Points*

The contexts of boundary interest points are similar to that of the region points. In order to avoid ill-defined boundaries, we put a few split points in the regions of the interest points, which defined semi-random split points. Fig. 3 shows that the image is separately divided by 3 random points and 3 semi-random points into partitions. The solid line cross the random points(marked as △), and the dotted line cross the semi random split points (marked as O). The test points marked as X are nearby the ground truth region which is marked as solid rectangle. All the neighbor points are marked as * shown in the image.

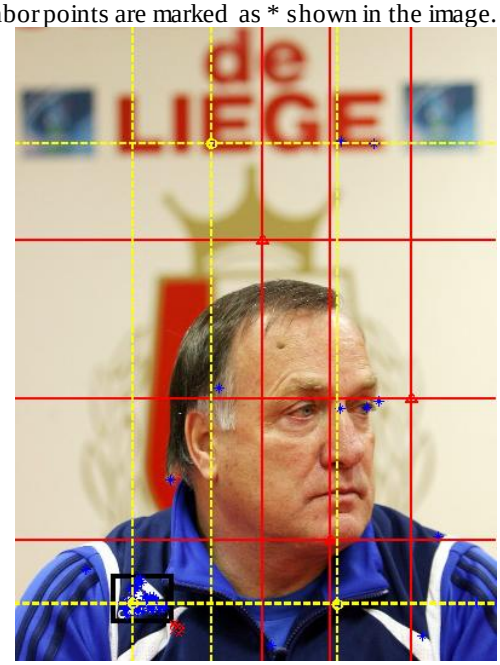

Figure 3. semi-random vs. random partition

The test points are not belonged to the ground truth regions. Therefore, the intersection area of their contexts regions and ground truth should smaller than those of the points in the ground truth region. We compare the semi-random partition with the random partition in searching Adidas logo in the Belgalogo dataset. In the experiment, the number of split points n=10, 15, 20 and 25. According to the neighbor points' distribution, all the neighbor points are clustered into m ($m \leqslant 6$) groups. m split points are selected in the m groups randomly. The rest points (n-m points) are selected randomly in the image.

Table 3 shows the average distance of the test points' context and ground truth regions of logo Adidas. The distance is cosine distance. The result shows the contexts of semi-random split points is more dissimilar than that of random ones.

TABLE III. THE AVERAGE DISTANCE OF TEST POINTS AND CENTER POINTS

| n | 10 | 15 | 20 | 25 |
|---|---|---|---|---|
| Random | 0.9269 | 0.9373 | 0.9484 | 0.9621 |
| Semi-random | 0.9375 | 0.9676 | 0.9638 | 0.9658 |

## V. CONCLUSIONS

Object searching in natural images is a challenging task in computer vision, due to lacking enough discriminative features. Our algorithm extracts enough interest points by LSH, computes context for every interest point to measure its discrimination, and clusters all interest points into serval groups to locate the bounding box for the target objects in each image. Our algorithm is test on the challenging Belgalogos dataset, the result shows it can improve features' discrimination effectively, and it is effective to locate more than one target objects in an image. But, because we locate the target objects by RANSAC algorithm, it is not good at locate objects when they are occluded or distorted greatly. Our future work includes how to locate target objects more accurately with more effective algorithm.

## REFERENCES


[1] J. Winn and N.Jojic. Locus: Learning object classes with unsupervised segmentation. In Proc. IEEE Intl. Conf. on Computer Vision, 2005

[2] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentation to discover objects and their extent in image collections.In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2006

[3] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2007.

[4] Y. Jiang, J. Meng, and J. Yuan, "Grid-based local feature bundling for efficient object search and localization," in Proc. IEEE Conf. Image Process., Sep. 2011, pp. 113–116

[5] S. Belongie, J. Malik, and J. Puzicha (April 2002). "Shape Matching and Object Recognition Using Shape Contexts". IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (24): 509–521.

[6] Xiaodong Yang, YingLi Tian. Texture representations using subspace embeddingsPattern Recognition Letters, Volume 34, Issue 10, 15 July 2013, Pages 1130-1137.

[7] Hong Lu, Wen-Lin Zou, Hong-Sheng Li, Yu Zhang, Shu-Min Fei.Edge and color contexts based object representation and tracking. Optik International Journal for Light and Electron Optics, Volume 126, Issue 1, January 2015, Pages 148-152.

[8] Fei-Fei Li; Perona, P. (2005). 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) . p. 524

[9] Lazebnik, S.; Schmid, C.; Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR'06) . p. 2169

[10] Chunjie Zhang, Xian Xiao, Junbiao Pang, Chao Liang, Yifan Zhang, Qingming Huang. Beyond visual word ambiguity: Weighted local feature encoding with governing region. Journal of Visual Communication and Image Representation, Volume 25, Issue 6, August 2014, Pages 1387-1398

[11] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6):381–395, 1981.

[12] Christoph H. Lampert, "Detecting objects in large image collections and videos by efficient subimage retrieval," inProc. IEEE Intl. Conf. on Computer Vision, 2009.

[13] Yuning Jiang, Jingjing Meng, and Junsong Yuan, Randomized Visual Phrases for Object Search, IEEE Computer Vision and Pattern Recognition (CVPR'12) 2012.

[14] M. Datar, N. Immorlica, P. Indyk, and V. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. DIMACS Workshop on Streaming Data Analysis and Mining, 2003.

[15 ]Alexis Joly and Olivier Buisson, Logo retrieval with a contrario visual query expansion, In Proceedings of the Seventeen ACM international Conference on Multimedia, 2009.