Combination of Key Point and Region-Based Recognition for Unsupervised Learning of Common Object in Multiple Images

Huiguang Yang
University of Illinois, Urbana
Urbana, IL
hyang30@illinois.edu

Abstract

A combination of key point-based and region-based recognition scheme is proposed in this study for unsupervised learning of common object from multiple images. The common object is a priori unknown in the image, and we need to simultaneously learn, detect and segment the common object from the image set. We make use of the robustness of key point-based scheme at the initial stage to learn and detect the common object in each image, and then apply region-based approach in the later stage to refine and complete the recognition as well as segment the object from the images. The saliency or common configuration for the common object in the images can be discovered by a two-dimensional Hough table, and the initial key point matches are verified and only those matches consistent with the common configuration will be retained. Several reference points (anchor points) on the common object in each image are computed based on the verified key points. After that the images are segmented and segment's spatial layout can be specified with respect to the anchor points in each image, which can help us to match the segments based on both shape and spatial layout. We apply the proposed scheme for detecting the common object from various scenes where the common object is a priori unknown. The performance of the scheme is validated by the experiments.

1. Introduction

Detecting the common object or object with common characteristics from various images is an important and interesting issue in image processing. To be specific, the problem is given a set of images, each contains a common but *unknown* object, we want to automatically identify this common object as well as detect it in each image. This task has a rich application background, for example, automatically detecting the same car or landmark in various images, detecting the same food in the different refrigerators, and much more.

Detecting common object in various scenes has a close connection with image registration, yet image registration is most often referring to the unification of the same or partially overlapped scene under different conditions (such as time, viewpoint etc.) [1][2], while our problem is trying to figure out the common object (but unknown *a priori*) appearing in various scenes. However either problem requires the steps of feature extraction and feature matching [1]. Feature is considered as the distinctive descriptor for representing local image characteristics, and commonly used features extracted from the image could be either point-based or region-based (sometimes even line-based) [1]. Consequently, the commonly used object recognition schemes would include key point-based approach and region-based approach.

The key point-based recognition approach is relying on the key point extracted from the image, where the key point could be corners [3-5], line intersections [6], high variance points [7], most distinctive points [8], or the SIFT key points [2][11] in the image. The key points are then matched across the images based on their local neighborhood property. For example in SIFT approach, the SIFT key point is extracted as the maxima and minima in the Difference-of-Gaussian image pyramid, and the key points are matched based on the gradient histogram extracted from the local neighborhood around each key point [11].

The key point-based approach is quite dominate in object recognition in last decade or two, because people consider the key point feature is more robust than the region-based feature or the contour-based feature. It is natural since the key point feature focuses on the micro-region of the image, its property is not easily affected by the occlusion, lighting change or distortion of object. Even though part of the object is occluded, the key point features extracted from the part which is not occluded will not be changed. Whereas in the region-based recognition scheme, if the object is partially occluded, the object segment property such as shape could be changed dramatically. Therefore the major advantage of key point-based recognition approach is robustness.

However key point-based approach also has its disadvantages. First of all we will lose the global view of the image when we use key point features, because we only

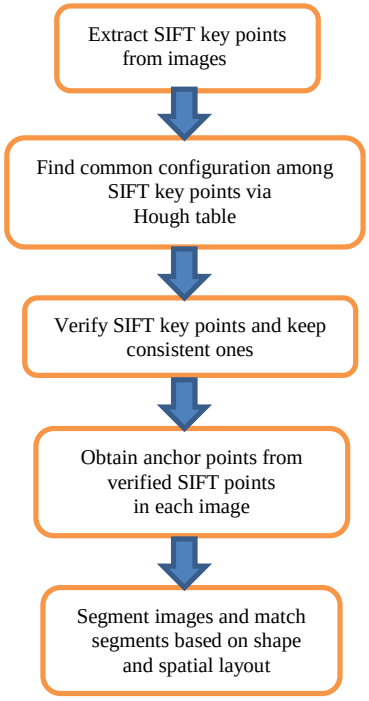

Figure 1. The flow chart of the approach.

focus on the tiny micro portion of the image and do not take into account any high level information. Because of this key point features may bring dramatic ambiguity in matching. Two key points match well does not mean the objects where they are detected could be similar in any sense, in fact purely key point matching can lead to huge amount of mismatches. The crucial remedy is to take into account the high level information in some other way, say in terms of spatial layout of the key points, to bring more constraints for the matching. Another drawback of the key point approach is that it cannot provide exact object boundary since it only matches points. We can only obtain the bounding box over the object if we know the object model and its bounding box.

In comparison with the key point-based recognition approach, the region-based recognition approach is relying on the image segmentation techniques [16-19], where the image is segmented into homogenous regions which are contrasted with the surroundings. Then the image segments are matched across the images, based on the segment properties such as shape, average intensity etc., as well as their spatial layout. Some segmentation algorithms segment the image in a hierarchical manner to obtain a segmentation tree [18], where one has a complete hierarchy of segments from fine to coarse scales, hence the hierarchical containment relationships among the segments and the adjacency relationships between the segments can also be explored and used in the segment matching.

In region-based recognition scheme object or segment shape plays a central role in recognition, since it is shift, scale and rotation-invariant. The region average intensity can be used as region feature as well, but it could change if the lighting condition in the image changes, while the region shape will not be affected by that. Therefore in region-based recognition the majority of the segment feature descriptors are focusing on the segment shape.

The advantage of region-based approach is that it can take into account high level information of the image, and hence reduce the ambiguity brought in by the features. For example, two segments with relatively sophisticated shape get matched is much more difficult than two similar key points get matched, therefore the matching is more reliable. At the same time, region-based approach can provide exact object boundary, so it can automatically detect and segment the object out of the image.

The disadvantage of region-based approach is that it is not very robust. The segment property can be easily affected by occlusion, lighting change (for example the shadow or highlight on the object will change the segmentation) and distortion. Also the segmentation algorithm itself can be unstable sometimes, the same object under different background can be segmented in somewhat different way, which brings big trouble for segment matching. The matching of segments can be very effective and distinctive if such match can be found, but in many cases no good match among the segments can be found across the images.

In viewing of the advantages and disadvantages of both schemes, here comes in the idea of combining the key point-based recognition approach with the region-based approach. We can make use of the robustness of key point-based scheme at the initial stage to learn the common object and detect it in images. Then we can compute several reference points on the object in each image based on the reliable key points, in order to specify any new key point or segment's spatial location. Finally we apply region-based recognition to obtain the image segments and match the segments based on the shape and spatial layout. Therefore we can detect and segment out the object from the images. The flow chart of the detailed approach is shown in Figure 1.

The structure of the paper is organized as follows. Methodology will be stated in Section 2, and Section 3 describes the segment features employed in region-based recognition in this study. Experimental results are provided in Section 4 and conclusion is given in Section 5.

2. Method

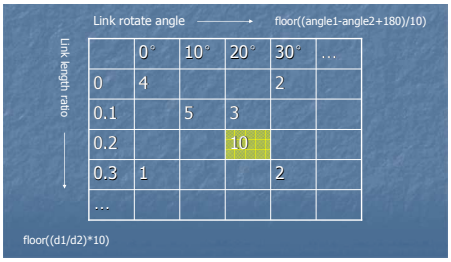

Figure 2. An example of the Hough table, where the two dimensions of the Hough table are link length ratio and link rotate angle. For each pair of original key points in the image we link them and find its correspondence in another image to compute the link length ratio and link rotate angle. We bin them in the Hough table and find the maximal value in the Hough table and therefore the common configuration.

Given a set of images, the first step is to extract SIFT key point from each of them. As is done in Lowe's paper [11], for each SIFT key point we find in image A, we look for its nearest neighbor among the SIFT key points we find in image B, meanwhile we check the distance ratio between the nearest neighbor and second nearest neighbor, if this ratio is below 0.8 we then accept this pair of key points as the original match. But even after such examination the obtained SIFT key point matches still contain huge amount of outliers and mismatches, we need to figure out a reliable subset among those matches representing the true correspondences for the common object. Since the ratio of inliers to outliners in the original matches is very low for the unsupervised case, typically less than 5-10 percent, finding out the real correspondences among them is quite challenging. Given we do not know the common object in advance and thus no object model, we need to discover the saliency or common configuration from the initial matches. We find the high dimensional Hough table is a very effective and convenient tool in achieving this goal. The important observation leading to this idea is that, if we have two key point pairs both are real matches, their links in both images should also form true correspondence, and their link length ratio and rotate angle between two images should be around a fixed value and be consistent with all the true matching link pairs between the two images, since all of them are on the common object. The common object may be scaled and rotated between the images, but such effect should be the same for all corresponding links. For now the off-plane rotation or the full affine transform is not considered yet, but the slight off-plane rotation can be well tolerated.

To create the Hough table (Figure 2), we link each pair of key points in image A and find its corresponding link in image B. We compute their length ratio and rotate angle

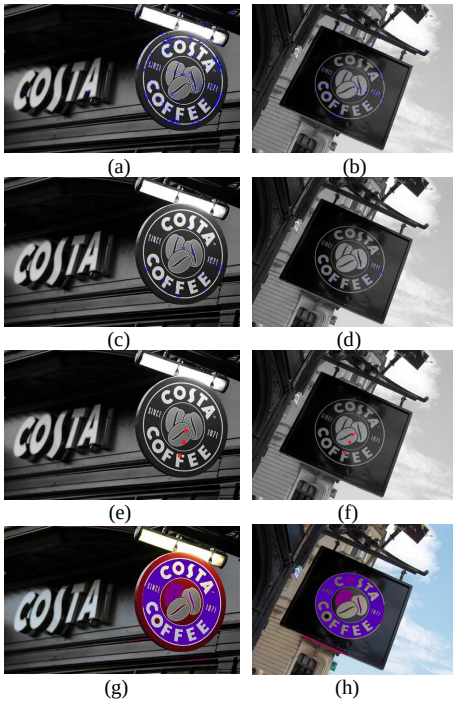

Figure 3. The illustration of the combined key point and region-based recognition approach. (a-b) original SIFT key points extracted in two images. (c-d) verified SIFT key points on the common object in two images. (e-f) the three anchor points obtained in each image. (g-h) the common object detected in two images, the same color denotes the corresponding regions (note the final matching result is still not yet perfect, I'm still working on it to make improvement).

between the two images and histogram them in the two dimensional Hough table with length ratio and rotate angle bins as two dimensions. We bin the length ratio in the step size of 0.1, from zero to the maximal length ratio in the image, normally this value is no larger than 10. The rotate angle has the bin size of 10 degrees, it goes from 0 to 360 degrees. All the real corresponding links have the length ratio and rotate angle around similar values, whereas the outlier key points may have these two values anywhere. Given the intuition that the object should appear in a compact region in the image, we only check the local neighborhood of certain size around each key point and the links therein. In this experiment we consider the neighborhood of the size 2/5 height and 2/5 width of the image around the key point.

After the Hough table is obtained, we find the maximal value in the Hough table and the corresponding values of length ratio and rotate angle are considered as the common configurations for the common object. With this common configuration we search for each key point in image A again, if the key point's local neighborhood contains

enough “correct links”, i.e. the links with parameters consistent with the common configurations, we then consider this key point (and its corresponding one in image B) as the true correspondence, otherwise the key point will be discarded. In this experiment we accept the key point with the “correct link” ratio in its neighborhood above half of the maximal “correct link” ratio in the image. This step is called the verification of the original SIFT key point.
After the verification of the key points, we should be able to obtain the key points only appearing within the object region and all form exact matches. Based on them, we can further compute three anchor points in each image by simply averaging the verified key points in certain region, for example the mean of all verified key points, the mean of first half and the mean of second half (then the three anchor points will be collinear, to avoid this we can use their mean, the mean of first 75% and the mean of latter 75%). The anchor points can serve as a reference in each image and any new key point or segment’s location can be determined based on them. The reason that three anchor points are used is because any new point’s location will be uniquely determined if we know its distances to three points in the 2D plane (if the three points are not collinear). With the help of anchor points we can specify each segment’s spatial layout with respect to them, which can help us in determining if two segments from two images form real match if their shapes are already similar.
After obtaining of anchor points in each image, the region-based recognition comes in at this stage. We segment the image using the segmentation-tree algorithm, hence we can obtain a complete hierarchy of the image segments from fine to coarse scale, which can help us to match segments in various levels and be more robust to the merging or splitting in segmentation. Also we can just segment the image region around the anchor points since that is where we believe the object lies. Then we match the segments in image A with the segments in image B in terms of similarity in both shape and spatial layout. In practice we blur the image with successive Gaussian kernels to obtain an image pyramid, and we segment each of the images in the pyramid using the segmentation-tree algorithm. In such a way we can obtain a “bag of segments” and will therefore have a better chance to match the corresponding segments which may vary slightly from image to image. With the matching of segments, we can find the exact boundary of the common object in each image and hence segment the object from the image. The illustration of the entire approach is shown in Figure 3.
For the image set containing multiple images, we can work on them pairwisely, i.e. we process pair of images at a time until we visit all the images. Or we can work on all the images altogether. In this case the Hough table will not be changed, we just have more elements from more images to be binned in the Hough table. Also since we only consider the key point link within each image, not the link across the images, the computation complexity scales linearly with the number of images.

3. The features in region-based recognition

In region-based recognition each segment is described by a bunch of features, in the hope that the more similar the segments are, the closer their features are in the feature space. In order to achieve size, location, orientation invariant recognition, the majority of the features will be focusing on the segment shape. The features extracted from the segment in this study include the moment of inertia, normalized perimeter, bounding circle area ratio, area over filled area ratio, boundary geometry, contour division and contour mapping vector.

3.1 Moment of inertia

In order to come up with a segment shape descriptor which is scale and rotation invariant, we propose a shape feature called the moment of inertia, inspiring by the same concept in physics. In physics, moment of inertia is a quantity to describe the “inertia” of object in the rotation, which is directly related with the object shape or the object mass distribution with respect to its mass center, and forms a good descriptor for shape.
For a given 2-D region, the moment of inertia is calculated as the vector summation of each region pixel’s squared distance to the mass center of the region.

$$I = \sum_{\vec{x} \in R} (\vec{x} - \vec{x}_C)^2 \qquad (1)$$

where R denotes the region and $\vec{x}_C$ denotes the location of mass center.
For the regions with the same shape but varying in size, it is easy to show that the moment of inertia goes with the fourth power of the region dimension scale, i.e. for the regions having the same shape but one is N times larger than the other (in length not in area), the moment of inertia is N^4 times larger (since the distance is N times larger and the total mass (or area) is N^2 times larger). Therefore we normalized the moment of inertia by the fourth power of region dimension to obtain the scale-invariant shape feature.

3.2 Normalized perimeter

Segment perimeter is also helpful in describing the segment shape. Yet since segments with the same shape may have

different sizes, we need to normalize the segment perimeter with the area.

$$\hat{L} = \sqrt{\frac{L^2}{S}} \quad (2)$$

where L is the segment perimeter and S is the area of the segment.

3.3 The bounding circle area ratio

Another shape descriptor is the ratio of segment area over its bounding circle area, which is a measure of the object compactness. In plus it is rotation-invariant – this is why the bounding circle is used instead of the bounding box.

3.4 Area to filled area ratio

Some segments have holes inside, hence we can measure segment's area to filled area ratio.

3.5 Boundary geometry

The boundary of the segment is certainly the key part to describe the segment shape. For a given segment we obtain the farthest distance ($d_{\max}$), nearest distance ($d_{\min}$) and mean distance (d_{mean}) from its boundary point to the mass center. We take the ratios of these quantities to obtain the features in describing the boundary geometry. Such features are scale invariant.

$$R_{\max/\min} = \frac{d_{\max}}{d_{\min}} \quad (3)$$

$$R_{\max/mean} = \frac{d_{\max}}{d_{mean}} \quad (4)$$

3.6 Contour division

The ultimate segment shape descriptor would rely on the contour of the segment. For many of above shape features, there is no one-to-one mapping between the feature and the shape. The same feature value can lead to different shapes. An appropriate contour descriptor can provide a much more accurate description for the shape. There has been several contour shape descriptors proposed in the literature such as the chain code representation [13] and the binary shape matrix [22] mentioned above. Here we propose another contour shape descriptor named contour division. Basically on the segment contour we find the points with the maximal/minimal distance to the segment mass center ($p_{\max}$, $p_{\min}$), respectively, and find the ratio of the length *along the contour* between these two points over the length of entire contour (we take the ratio which is less than 0.5). Although contour division would not specify the shape uniquely, it offers valuable information about the segment boundary.

$$\rho = \min\left(\int_{p_{\max}}^{p_{\min}} ds \Big/ \int_C ds,\ \ 1 - \int_{p_{\max}}^{p_{\min}} ds \Big/ \int_C ds \right) \quad (5)$$

3.7 Contour mapping vector

In order to describe the segment contour more accurately, we further propose a contour shape descriptor called contour mapping vector. In essence we pick up N equally distributed points on the segment contour (note this is equally distributed points along the contour length, not the angle division) and compute their distances to the segment mass center and store the N distances in a vector. The N distances are then normalized by the mean distance between the segment contour and mass center to achieve scale invariant. To achieve rotation invariant, we locate the point with the maximal distance to the mass center on the contour and start the N points from there. When N is large this feature can describe contour very accurately. Meanwhile it is location, scale, rotation invariant. The ith element of the contour mapping vector is

$$v_i = \frac{dist(p_i, p_C)}{\bar{l}_C}, \quad i = 1,2,\cdots,N \quad (6)$$

where $p_1 = p_{\max}$ is the point with the maximal distance to the mass center on the contour, p_i is the ith equally distributed point on the contour started from p_1 , p_C is segment mass center, $\bar{l}_C$ is the mean distance between the contour and the mass center.

4. Experiments

We carry out experiment on a set of images containing a handcrafted bottle, where the object size, lighting condition, view angle may vary from image to image. We do not have the prior knowledge on what is inside the images, and we want to recognize the common object as well as detect and segment it out of each image. The original SIFT key points are first extracted for each image as is shown in Figure 4 (a-b). The verified SIFT key points are obtained via the Hough table binning as is shown in Figure 4 (c-d). Three anchor points are computed in each image as is shown in

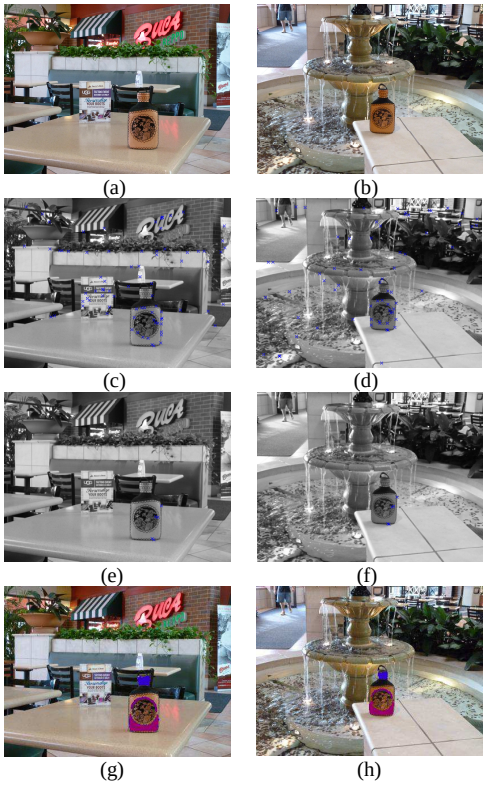

Figure 4. The experiment on the handcrafted bottle. (a-b) original images. (c-d) original SIFT key points detected in two images. (e-f) verified SIFT key points on the common object in two images. (g-h) the common object detected in two images, the same color denotes the corresponding regions.

Figure 4 (e-f). Then the images are segmented and the segments are matched based on their shape and spatial layout across the images as is shown in Figure 4 (g-h).

5. Conclusion

The combination of key point-based and region-based recognition approach can take the advantages of both schemes – the robustness of the key point-based approach and the high-level information as well as the exact object boundary obtained in the region-based approach. Combining the two approaches in a proper manner will certainly help the recognition. We find the Hough table is very effective and convenient in discovering the saliency in the images, especially when the ratio of inliers to outliers is low. Once the location of the common object in each image is detected, the anchor points obtained on the object will offer reference for other segments or key points in order to determine their spatial layout in the image, and hence greatly facilitate segment matching.
Right now the scheme is still not yet perfect, due to the segmentation instability (the same object may not be segmented in the same way in different images) many visually similar segments cannot get matched in the final result, and not every part of the common object can be matched and detected. Therefore the common object boundary may be incomplete. I'm still working on these issues to further improve the approach.

References

[1] B. Zitova and J. Flusser, Image registration methods: a survey, Image and Vision Computing, 21, 977–1000, 2003.
[2] Y. Zhu, S. Cheng, V. Stankovic, L. Stankovic, Image registration using BP-SIFT, J. Vis. Commun. Image R., 24, 448–457, 2013.
[3] D. Bhattacharya, S. Sinha, Invariance of stereo images via theory of complex moments, Pattern Recognition, 30, 1373–1386, 1997.
[4] Y.C. Hsieh, D.M. McKeown, F.P. Perlant, Performance evaluation of scene registration and stereo matching for cartographic feature extraction, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14, 214–237, 1992.
[5] C.Y. Wang, H. Sun, S. Yadas, A. Rosenfeld, Some experiments in relaxation image matching using corner features, Pattern Recognition, 16, 167–182, 1983.
[6] A.S. Vasileisky, B. Zhukov, M. Berger, Automated image coregistration based on linear feature recognition, Proceedings of the Second Conference Fusion of Earth Data, Sophia Antipolis, France, 59–66, 1998.
[7] M. Ehlers, Region-based matching for image registration in remote sensing databases, Proceedings of the International Geoscience and Remote Sensing Symposium IGARSS'91, Espoo, Finland, 2231–2234, 1991.
[8] B. Likar, F. Pernus, Automatic extraction of corresponding points for the registration of medical images, Medical Physics, 26, 1678–1686, 1999.
[9] Z. Zheng, H. Wang, E.K. Teoh, Analysis of gray level corner detection, Pattern Recognition Letters, 20, 149–162, 1999.
[10] K. Rohr, Landmark-Based Image Analysis: Using Geometric and Intensity Models, Computational Imaging and Vision Series, vol. 21, Kluwer Academic Publishers, Dordrecht, 2001.
[11] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, 60 (2), 91–110, 2004.
[12] W.H. Wang, Y.C. Chen, Image registration by control points pairing using the invariant properties of line segments, Pattern Recognition Letters, 18, 269–281, 1997.
[13] H. Li, B.S. Manjunath, S.K. Mitra, A contour-based approach to multisensor image registration, IEEE Transactions on Image Processing, 4, 320–334, 1995.
[14] J. Canny, A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 679–698, 1986.

[15] D. Marr, E. Hildreth, Theory of edge detection, Proceedings of the Royal Society of London, B 207, 187–217, 1980.
[16] A. Goshtasby, G.C. Stockman, C.V. Page, A region-based approach to digital image registration with subpixel accuracy, IEEE Transactions on Geoscience and Remote Sensing, 24, 390–399, 1986.
[17] N.R. Pal, S.K. Pal, A review on image segmentation techniques, Pattern Recognition, 26, 1277–1294, 1993.
[18] E. Akbas and N. Ahuja, From ramp discontinuities to segmentation tree, Asian Conference on Computer Vision (ACCV), 2009.
[19] N. Ahuja, A transform for multiscale image segmentation by integrated edge and region detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, 18, No. 12, 1211-1235, 1996.
[20] S. Abdelsayed, D. Ionescu, D. Goodenough, Matching and registration method for remote sensing images, Proceedings of the International Geoscience and Remote Sensing Symposium IGARSS'95, Florence, Italy, 1029–1031, 1995.
[21] T. Peli, An algorithm for recognition and localization of rotated and scaled objects, Proceedings of the IEEE, 69, 483–485, 1981.
[22] A. Goshtasby, Description and discrimination of planar shapes using shape matrices, IEEE Transactions on Pattern Analysis and Machine Intelligence, 7, 738–743, 1985.
[23] D. Skea, I. Barrodale, R. Kuwahara, R. Poeckert, A control point matching algorithm, Pattern Recognition, 26, 269–276, 1993.
[24] M.K. Hu, Visual pattern recognition by moment invariants, IRE Transactions on Information Theory, 8, 179–187, 1962.
[25] X. Dai, S. Khorram, A feature-based image registration algorithm using improved chain-code representation combined with invariant moments, IEEE Transactions on Geoscience and Remote Sensing, 37, 2351–2362, 1999.
[26] A. Goshtasby, G.C. Stockman, Point pattern matching using convex hull edges, IEEE Transactions on Systems, Man and Cybernetics, 15, 631–637, 1985.