Ranking the Local Invariant Features for the Robust Visual Saliencies

Shengping Xia [1, 2] Peng Ren [2] Edwin R. Hancock [2]
[1] ATR Lab, School of Electronic Science and Engineering,
National University of Defense Technology, Changsha, Hunan, P.R.China 410073
[2] Department of Computer Science, University of York, York YO1 5DD, UK
xiasp1227@vip.sohu.com pengren@cs.york.ac.uk erh@cs.york.ac.uk

Abstract

Local invariant feature based methods have been proven to be effective in computer vision for object recognition and learning. But for an image, the number of points detected and to be matched may be very large, or even redundantly represent the shape information present. Since selective attention is a basic mechanism of the visual system, we explore whether there is a subset of salient points that can be robustly detected and matched.

We propose a method to rank the redundant local invariant features. The results prove that the top ranked points capture the salient information effectively. The method can be used as a pre-processing step for the Bag-of-Feature based methods or graph based methods. Here they simplify the complexity of the processes, such as training, matching and tracking.

1 INTRODUCTION

Selective attention is the ubiquitous mechanism that regulates the bottleneck between the massively-parallel world of sensation and the serial world of cognition [17]. The first step is low-level visual saliency selection. Intuitively, it is possible that the initial selection may not be a good indicator of the subjectively interesting scene elements. Observers may discard the majority of features as uninteresting, relying instead on different mechanisms to isolate more subjectively interesting locations [17].

An important question is the order of the selection. A visual saliency map model, in which the salient points are ranked for visual selection, is developed in [14]. Combining the saliency map model and the database of LabelMe [12] , they found that the first step in locating interesting elements in a scene is largely constrained by low-level visual properties and is a strong predictor of which regions might be interesting to human observers [4].

Recently local invariant feature extraction[8][10][11] (*LIFE*) methods, such as *SIFT*, *PCA-SIFT*, *GLOH* and *SURF*, have been proven successful and widely used for object classification [2][3][9], scene classification[1][5][6], and video retrieval[7][13]. A common draw back of *LIFE* methods is that too many orderless feature points are extracted from each image. This results in two problems, one is the exhaustive searching for all of the features, and the other is that too many false positives are produced. The feature selection problem has been widely researched for the Bag-of-Features methods [6][7][18] , using a holistic database to select a set of representative features. The method may be regarded as a combination of top-down and bottom-up methods to locate the best application-driven features, while not directly involving the exhaustive search or the false positive problem.

In this paper, we will focus on using bottom-up method to decrease the number of the features for the following processing stages. The key idea is that since *LIFE*- points are invariant to scale, illumination, rotation, and a small range of view point change, if we actively adjust these imaging conditions, the robust invariant features should be retained or otherwise eliminated. Motivated by this, we propose a method to rank the local invariant features with a correspondence to their visual saliency.

2. Review of the *SIFT* Algorithm

According to a recent comprehensive test result on large scale database in [9] , *SIFT* still appears to be the most appealing descriptor for practical uses, and has become a standard of comparison. However, in order to achieve the best possible performance, it is necessary to use a combination of several detectors and descriptors [9], which will incorporate multi-complementary features. For simple description, we simply experiment with *SIFT* in the following. However we emphasize that our methods could be

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

adapted to alternative families of *LIFE*-methods or any combinations of them.

By focusing on the most salient, repeatable image structures, *LIFE* methods normally incorporate three basic steps: detection, description and matching.

For *SIFT*, the detection step is designed to locate image regions that are salient not only spatially but also across different scales. Candidate locations are initially selected from local extrema in difference of Gaussian (DoG) filtered images in scale space [3].

As a second stage, to obtain the descriptor, a histogram of gradient directions in a patch sampled around the interest point is used as an estimate of orientation. In order to achieve orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the keypoint orientation. The size of this patch is determined by the scale of the corresponding extremum in the DoG scale space. This makes the descriptor scale invariant and this is an important property of the representation. The feature descriptor consists of histograms of gradient directions computed over a 4x4 spatial grid. The gradient directions are quantized into eight bins so that the final feature vector has dimension 128 (4x4x8). Finally, the feature vector is unified to reduce the effects of illumination change according to the method proposed in [3].

Here we use a vector $\vec{X}^t = (x_1^t, x_2^t)'$ to describe the coordinates of each *SIFT* point in the corresponding initial image, a vector $\vec{R}^t = (r^t, \alpha^t)$ corresponding to the gradient magnitudes and orientations at $\vec{X}^t$, a vector $\vec{U}^t = [u_1^t, u_2^t, \cdots, u_n^t]^T$ representing the local *SIFT* descriptors, where n is 128, and t is the index of this feature point. For short these can be denoted using the set:

$$V^t = \{\vec{X}^t, \vec{R}^t, \vec{U}^t\},\ \vec{V} = \{V^t, t = 1,2,..., m\} \qquad (1)$$

For matching, the local features play the role of “visual words” in *Bag-of-Feature* models [6], and the nodes in graph based models [5]. The number of local features within these models has an exponential effect on the learning time, despite the use of efficient search methods. Clearly, more features give more coverage of the objects, but it makes the model slower to learn and introduces over-fitting problems due to the increased numbers of parameters [6][7][18].

3. Ranking the robust salient features

The method proposed in this paper, for which a schematic diagram is given in Figure1, uses an active mechanism to locate the robust invariant feature points. For static images, a series of images are generated with some random perturbations. For 3-D objects, we actively change view points, view distance, lightening conditions and background. We take the videos and the generated images in the same manner. We simulate the process for ranking the robust salient features in three ways: digital simulation, semi-simulation and ground truth data driven.

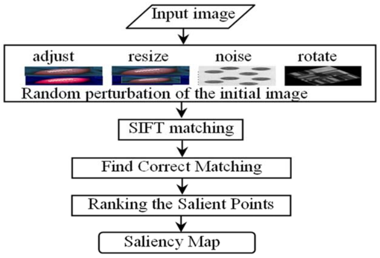


Figure 1 Diagram of ranking the salient features

3.1 Digital simulation method

We input an initial image to the image synthesis system, $\vec{V}_{ini}$ is obtained and the V^t_{ini} might be ordered according to the gradient magnitudes of $\vec{R}^t_{ini}$. Corresponding to the watch process, we generated some new images according to a Monte Carlo method, adapting the orthogonal experimental designation method.

We take four factors into account including image readjusting, image resize, and image noise and image rotation. A group of images can be generated randomly, which might be regarded as the changes of different imaging conditions.

These images are then SIFT-ed and matched with the initial image. The fraction of the correct matching (*FCM*) for every SIFT points is η^t, and the gradient magnitudes of $\vec{R}^t = (r^t, \alpha^t)$ is unified according to the maximum-minimum transform in the group of $\vec{R}_{ini}$, which is denoted as ζ^t. As a measure in our method the measurement of the ranking of the saliency is $\rho^t = \eta^t * \zeta^t$.

3.2 Semi-simulation method

In this case, the most significant difference with Section 3.1 is the source of the images. We first displayed the images on the screen or printed them on a paper. Then a Logitech camera (25*FPS, 320×240 image size)* were used to acquiring videos corresponding to the images under normal office conditions, where the camera was moved around the image, varying in distances and sometimes with rotation.

3.3 Ground truth data driven method

In Section 3.2, because the images are 2-D, which are very different from the ground truth 3-D objects, we developed further experiments for 3-D objects. The Logitech camera is also used to acquire the video. The schematic diagram is the same to the one in Figure 1. However in the process of Find-Correct-Matching, since we do not know the correctness of the *SIFT* matching result, which is different from Section 3.1, it is hard to count the correct correspondences. We have developed a spectral graph and geometric registration combined method to locate the correct matching under the similarity of the vector of spectral graph.

4. Experimental results and discussion

4.1 Experimental Results

In Figure 2, the images in the first column are the initial images, the second column are the ranked *SIFT* points (deep blue circles) overlaid on the initial images (the red lines are the edges of the Delaunay graph of these points), and the third column are the whole *SIFT* points marked on the initial images.

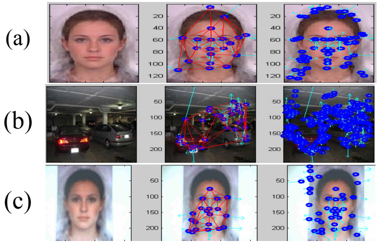


Figure 2 Ranked salient features and images

In Figure 2(a), 20 of the 64 points overlaid on the initial image, the first 10 points locate on the face, corresponding to the mouth, the eyes, the nose, the face, the jowl and the hair. Most of the points are salient enough to attract the visual attention.

In Figure 2(b), 30 of the 334 points displayed, the first 30 points cover the car lamps, the brightest ceiling lamps, and the center of the big dark area. The points are the most distinctive in the scene. More examples are shown in Figure 3.

In Figure 2(c), the images were frames of a video taken by a camera pointing to the image in Figure 2(a) displayed on a screen. Most of the salient points are in the ranked highly, as shown in Figure 2(c), which closely similar to Figure 2(a).

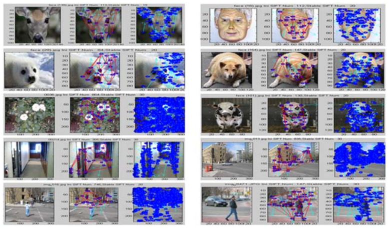


Figure 3 More Examples of experimental results

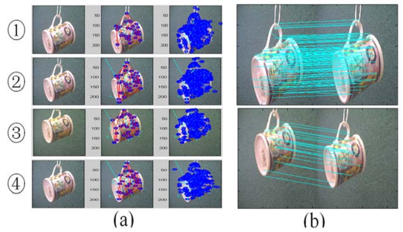


Figure 4 Ranking from ground truth data and matching

In Figure 4 (a), the results in ①, ②, ③ and ④ correspond to 4 different short sequences of video (30 frames) of the cup. Four groups of different ranked salient points are obtained, for which the points ranked in the top 30 are displayed. There are 20 common salient points in the four groups, an average 26 points matched between these 4 groups (Figure 4(b) bottom), and there are 41 different common points. While for about 300 *SIFT* points in each frames, there are over 130 points matched according to the initial *SIFT* matching. The correctness of the matched is unknown and hard to verify, as show in Figure 4(b) (top).

4.2 Evaluation

The COIL-100 data-set is used to evaluate the effectiveness of the method. At first, the images of camera angle 0°, 5° or 10° for 50 objects are selected to be processed with the proposed method. A part of 30 ranked feature points for each image were obtained, which construct a ranked *SIFT* point data-set, denoted as *R-SiftDB*. Secondly, we select 50 images, whose angle is 355° or 20°, for the corresponding objects. For every image, the first 20 ranked *SIFT* points for every image is selected for matching with the *R-SiftDB*. For each feature points ideally, there should be 3 points exactly matched. The most points and the corresponding images are shown in Figure 5. The order 1, 2, 3, to 9, is the order of matching similarity of the feature points. We manually reviewed the correctness of the matching.

In Figure 6(a), x-axis is the rank order of the feature points, and y-axis is the fraction of the correctly matching. The * is the *FCM* corresponding to position 1 of Figure 5, the O corresponds to position 2, and the □ corresponds to the position 3. In Figure 6(b), the + corresponds to the *FCM* when any one of the matched features in the first three position of Figure 5 is correctly matched (denoted as Case1-of-3). And the O corresponds to the *FCM* when at least 2 of the first three matched features are correctly matched (denoted as Case 2-of-3).

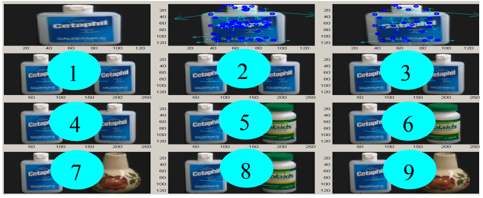
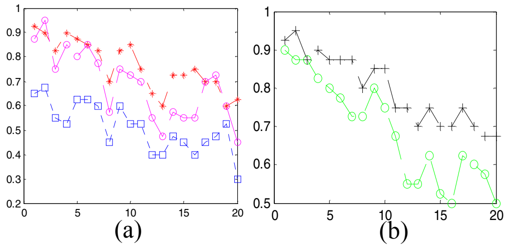


Figure 5 9 most matched points displayed



(a) (b)

Figure 6 *FCM* of the first 20 features

From Figure 6, the F*CM* decreases with the order of the rank. But the average *FCM* of the first 20 points is close to 70% in Case 2-of-3 and the average *FCM is about* 80%. For normal *SIFT* matching, it is very hard to manually review the correctness of the matching for the orderless mass of *SIFT* points. However a rough estimate of the average *FCM* is no more than 50%, which is also proved in [10].

This proves that not all *SIFT* points are actually invariant. Based on our method, a group of most robust, invariant and salient feature points can be extracted for further modeling.

5. Conclusions

In this paper we proposed a method to rank the local invariant features which are obtained by *SIFT* of which is easy extend to other *LIFE*-methods.

The rank index of the invariant features is of some distinct correspondence to the points which are of visual saliency. It is a direct way to select the robust and salient features according to the rank, which will drastically decrease the complexity of the following process, such as the training of the *Bag-of-Features* model, the matching of the graph based object recognition model and the moving object tracking of the template based methods, etc.

References

[1] Anna Bosch, etc. Scene Classification Using a Hybrid Generative/ Discriminative Approach. *IEEE Trans. PAMI*, 30(4):1-16, APRIL 2008.

[2] Csurka, G., etc. Visual categorization with bags of key points. In *ECCV Workshop on Statistical Learning in Computer Vision*. 1-22, 2004.

[3] D. G. Lowe. Distinctive image features from scale-invariant key points. *IJCV*, 60(2):91–110, 2004.

[4] Elazary, L., & Itti, L. Interesting objects are visually salient. *Journal of Vision*. 8(3):1–15, March 2008.

[5] Erik B. S., etc. Describing Visual Scenes Using Transformed Objects and Parts. *IJCV*, 77: 291–330, March 2008.

[6] Fei-Fei, L. etc. A Bayesian hierarchical model for learning natural scene categories. *CVPR*, 2:524–531, 2005.

[7] Fergus, R., etc. Learning object categories from Google's image search. *ICCV*, 2:1816–1823, 2005.

[8] Herbert Bay, etc. SURF: Speeded Up Robust Features. *ECCV*, Part I, LNCS 3951: 404–417, 2006.

[9] J. Zhang, etc. Local Features and Kernels for Classification of Texture and Object Categories. *IJCV*. 2(73):213-238, 2007.

[10] Ke, Y.etc. PCA-SIFT: A more distinctive representation for local image descriptors. *CVPR*, (2): 506-513, 2004.

[11] Mikolajczyk, etc. Scale and affine invariant interest point detectors. *IJCV*, 60:63-86, 2004.

[12] Russell, B. C., Torralba, A., etc. LabelMe: A database and Web-based tool for image annotation. *MIT AI Lab Memo AIM-2005-025*. http://labelme.csail.mit.edu/.

[13] Sivic, J. etc. VideoGoogle: A text retrieval approach to object matching in videos. *ICCV*, 2:1470–1477, 2003.

[14] Itti, L., Koch, C., & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20, 1254–1259, 1998.

[15] Itti, L. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12, 1093–1123, 2005.

[16] Einhäuser, W., etc. A bottom–up model of spatial attention predicts human error patterns in rapid scene recognition. Journal of Vision, 7(10):6, 1-13, (2007)

[17] Treue, S. Visual attention: The where, what, how and why of saliency. *Current Opinion in Neurobiology*, 13(4), 428–432. 2003.

[18] Svetlana Lazebnik, etc. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *IEEE CVPR*, 2169-2178, 2006.