Biologically Inspired Scene Context for Object Detection Using a Single Instance

Changxin Gao*, Nong Sang, Rui Huang

Key Laboratory of Ministry of Education for Image Processing and Intelligent Control, School of Automation, Huazhong University of Science and Technology, Wuhan, China

Abstract

This paper presents a novel object detection method using a single instance from the object category. Our method uses biologically inspired global scene context criteria to check whether every individual location of the image can be naturally replaced by the query instance, which indicates whether there is a similar object at this location. Different from the traditional detection methods that only look at individual locations for the desired objects, our method evaluates the consistency of the entire scene. It is therefore robust to large intra-class variations, occlusions, a minor variety of poses, low-revolution conditions, background clutter etc., and there is no off-line training. The experimental results on four datasets and two video sequences clearly show the superior robustness of the proposed method, suggesting that global scene context is important for visual detection/localization.

Citation: Gao C, Sang N, Huang R (2014) Biologically Inspired Scene Context for Object Detection Using a Single Instance. PLoS ONE 9(5): e98447. doi:10.1371/journal.pone.0098447

Editor: Paul Graham, Huazhong University of Science and Technology, China

Received October 18, 2013; **Accepted** May 4, 2014; **Published** May 28, 2014

Funding: This research is supported by the Project of the National Natural Science Foundation of China (No. 61105014), and the Fundamental Research Funds for the Central Universities (No. HUST:2013TS132). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cgao@hust.edu.cn

Introduction

Visual detection/localization is one of the basic functions of human vision. Most object detection methods are learning-based, which use lots of positive and negative samples to train a classifier. Training is often a time-consuming process and the training samples are very difficult to collect and manually label in many domains. Training-free image analysis methods are becoming more and more popular in recent years [1,2]. This paper addresses the problem of object detection in the training-free setting, namely, object category detection using only a single instance.

As a classical training-free computational model, template matching is very effective for object localization, by computing point-to-point correlation of a model pattern with the image pattern. It has been widely used in object detection, object tracking, feature detection, and motion estimation, etc. However, this technique is designed for object *instance* detection, i.e., searching only for the same instance as the template, because single templates are inadequate for encoding large intra-class variations, and local appearance features have limited robustness to occlusion, pose, low-resolution conditions and background clutter. The human visual system is remarkably robust and can work effectively even under highly uncertain and ambiguous conditions, where contextual information plays a critical role [3–5]. This paper focuses on improving the single instance based object detection method using contextual information.

Given the excellent performance of human vision in this task, it is reasonable to look to biology for inspiration. After surveying the relevant psychophysical and neurophysiological research, we find that *scene-centered viewpoint* [6–8] and *change blindness* phenomenon [9,10] may provide help to this task. Scene-centered viewpoint suggests that humans can capture the abstract meaning of a real world scene only using their global spatial layout without necessarily knowing the objects in the scene. Change blindness suggests that human often fail to notice the relatively large changes to visual scenes.

Scene-centered viewpoint

Traditional conceptions of research in computer vision treat objects as the atoms of recognition. In contrast, some experimental studies have suggested that the recognition of real world scenes may be initiated from the encoding of the global configuration, without necessarily identifying the objects they contain [4,7,11,12]. Humans are able to comprehend the amount of perceptual and semantic information (the semantic category of each image as well as a few objects and their attributes) in less than 200 milliseconds, which is referred to as getting the gist of the scene [8,13,14]. Biederman has demonstrated human identify a scene as quickly and as accurately as a single object [15]. The semantic category of most real world scenes can be inferred from their spatial layouts [6,12,16,17,18]. These experiments in scene perception and visual search suggest that the visual system first processes contextual information in order to index object properties. Following these results, some computer vision methods recognize an image using global context representation, termed as context-based approaches.

Change blindness

Humans often fail to notice the large changes to visual scenes, called change blindness. Several studies on change blindness have demonstrated that subjects can be totally blind to object changes (such as displacement and suppression) even when these changes

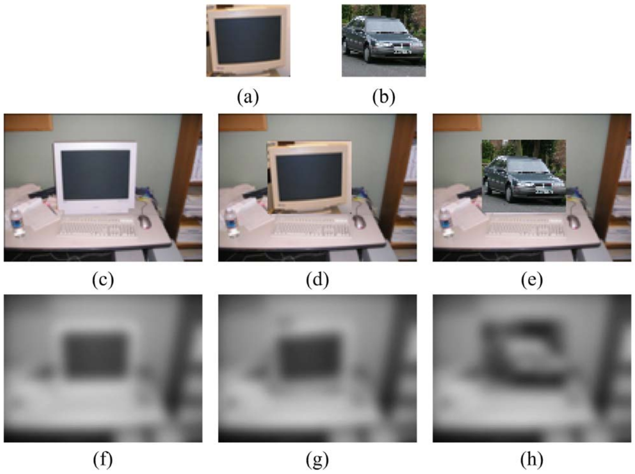


Figure 1. Illustration of the variety of coarse scene layout (produced using Gaussian blur with $\sigma = 4.0$) by replacing object instances in the scene with templates (another object instance). All the scene images are from the LabelMe database [25]. (a) and (b) are monitor and car templates respectively; (c) shows a scene which contains a monitor; (f) shows the layout of scene shown in (c); (d) and (e) show the results after replacing the monitor in the scene with the two templates; (g) and (h) show the scene layouts corresponding to images with replaced instances.
doi:10.1371/journal.pone.0098447.g001

affect meaningful parts of the scene [18]: observers frequently fail to notice when the central actor in a simple motion picture is replaced by a different person wearing different clothing and when the person they are talking to is surreptitiously replaced by a different person [10,19]; observers successfully recognize a previously attended object on a memory test even when they have failed to detect a change to that object [20]; observers can remain the pre- and post-change information at better than chance levels, even when they have failed to detect the change [21]. These studies suggest that human representations of visual scenes are sparse and incomplete [9,22]. It should be noted that the "large changes" in these studies indicate that objects are replaced with another object instance from the same category.

The motivation inferred from these two biological findings is that the spatial configuration of a scene is very important to human visual perception, and sufficiently big changes will break the structure of the scene while smaller changes will not. The change in global spatial layout is imperceptible when an object instance in a scene is replaced by another instance (at the same scale) from the same category, while it is more pronounced when replacement happens between distinct categories, as shown in Fig. 1. We use the blurred images as the coarse layout of the scenes. Despite the monitor template contains a small amount of background contents and the resulting images Fig. 1(c) and Fig. 1(d) are noticeably different, the layouts are very similar. The correlation coefficient (gray values are used) between Fig. 1(f) and

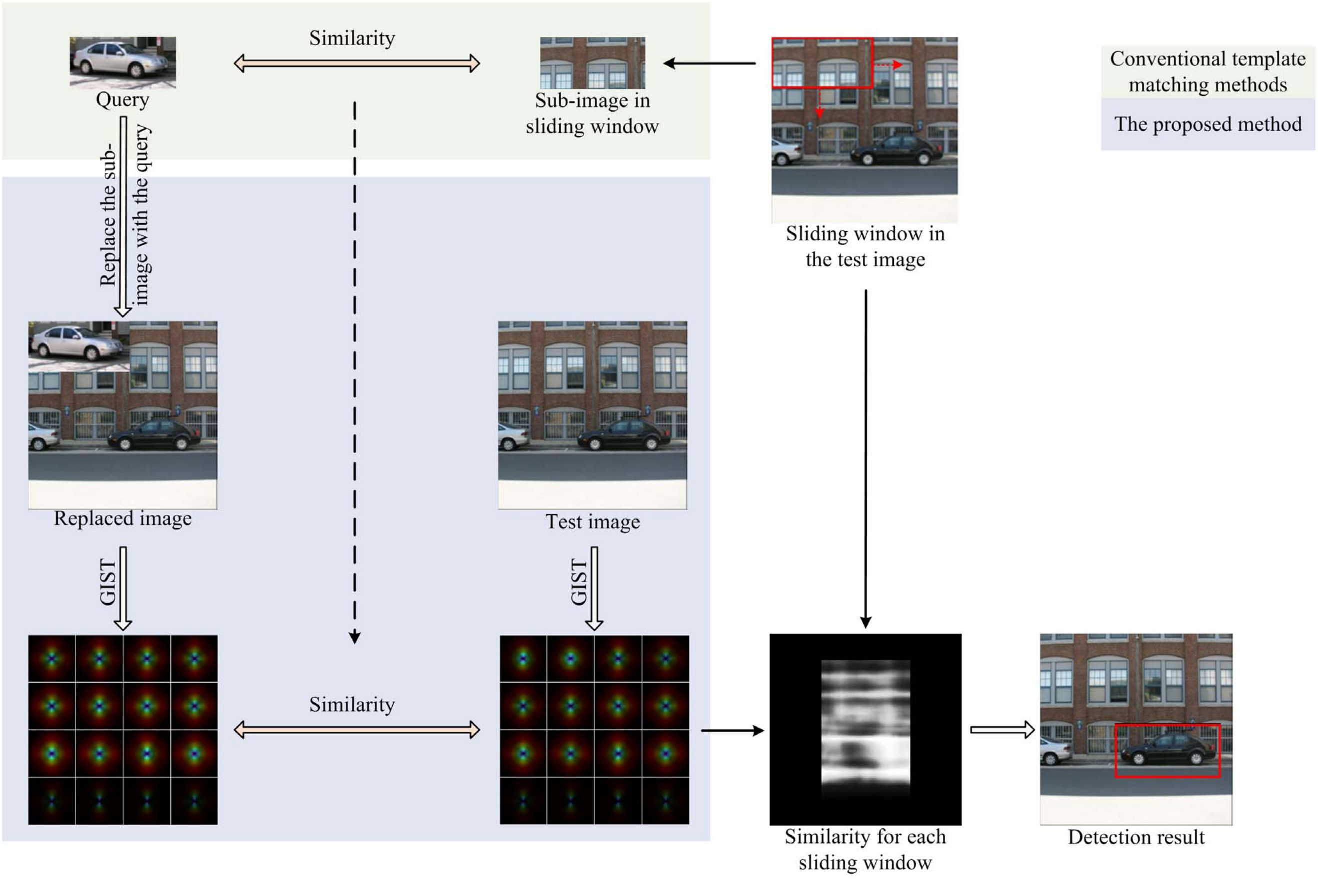
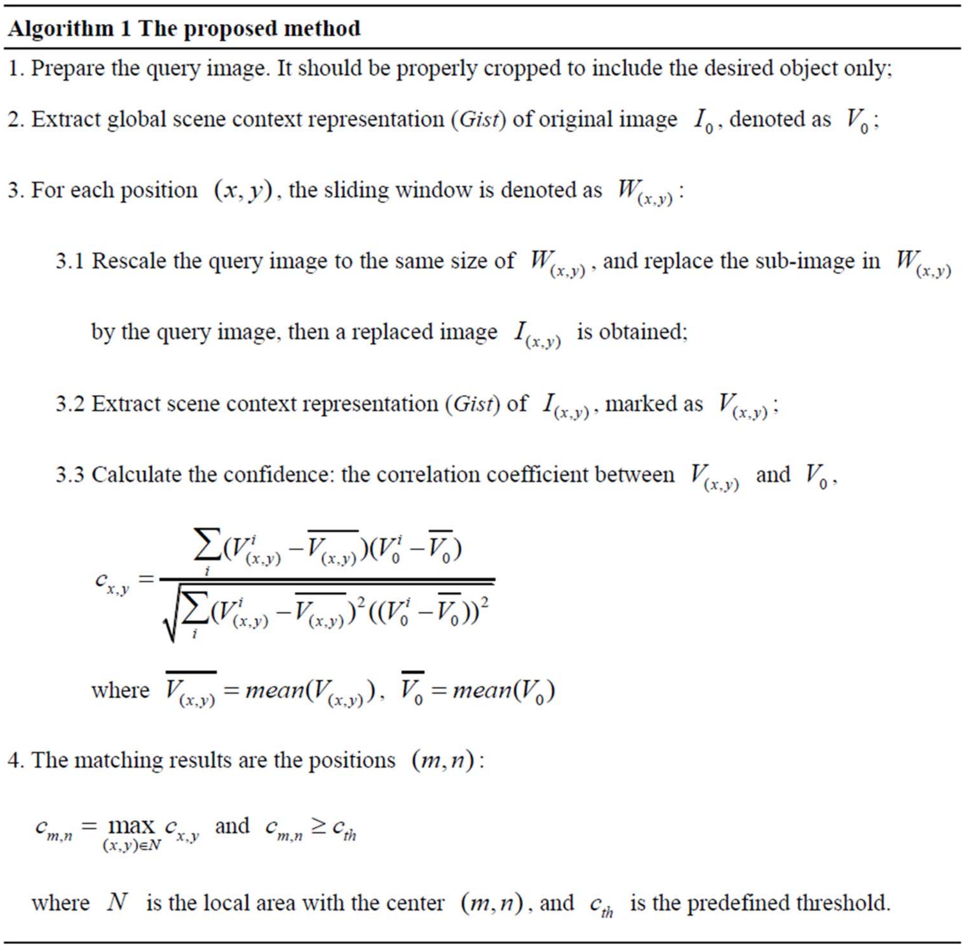

Similarity
Query
Sub-image in sliding window
Conventional template matching methods
The proposed method
Replace the sub-image with the query
Sliding window in the test image
Replaced image
Test image
GIST
GIST
Similarity
Similarity for each sliding window
Detection result

Figure 2. Sketch of the difference between conventional template matching and our template matching. Conventional template matching methods measure the similarity of the template and the sub-window; while our template matching method measures the similarity of the scene image; and the new scene image with the replaced window. All the scene images are from the LabelMe database [25], and the template is cropped from an image also in the LabelMe dataset.
doi:10.1371/journal.pone.0098447.g002



```
Algorithm 1 The proposed method
1. Prepare the query image. It should be properly cropped to include the desired object only;
2. Extract global scene context representation (Gist) of original image $I_0$, denoted as $V_0$;
3. For each position $(x, y)$, the sliding window is denoted as $W_{(x,y)}$:
    3.1 Rescale the query image to the same size of $W_{(x,y)}$, and replace the sub-image in $W_{(x,y)}$
        by the query image, then a replaced image $I_{(x,y)}$ is obtained;
    3.2 Extract scene context representation (Gist) of $I_{(x,y)}$, marked as $V_{(x,y)}$;
    3.3 Calculate the confidence: the correlation coefficient between $V_{(x,y)}$ and $V_0$,
        $c_{x,y} = \frac{\sum_i (V^i_{(x,y)} - \overline{V_{(x,y)}})(V^i_0 - \overline{V_0})}{\sqrt{\sum_i (V^i_{(x,y)} - \overline{V_{(x,y)}})^2 ((V^i_0 - \overline{V_0}))^2}}$
        where $\overline{V_{(x,y)}} = mean(V_{(x,y)})$, $\overline{V_0} = mean(V_0)$
4. The matching results are the positions $(m, n)$:
    $c_{m,n} = \max_{(x,y)\in N} c_{x,y}$ and $c_{m,n} \geq c_{th}$
    where $N$ is the local area with the center $(m, n)$, and $c_{th}$ is the predefined threshold.
```

Figure 3. The flowchart of the proposed algorithm.
doi:10.1371/journal.pone.0098447.g003

Fig. 1(g) is 0.98 and the one between Fig. 1(f) and Fig. 1(h) is 0.76. It illustrates that the change in global spatial layout is imperceptible when the replacement happens between the instances within the same category, and it is sensitive to those replacements between distinct categories.

To this end, we present a single instance based object category detection method using global scene context, which uses the constraints of scene context to improve the performance of matching (Preliminary version of portions of this work has been published in [23]). This method is quite different from conventional template matching methods: conventional methods measure the similarity of a query template and a sub-image using local features within them, while our method measures the compatibility of the two by replacing the sub-image with the template and evaluating the change in the scene context. The differences between the two methods are illustrated in Fig. 2.

To achieve this goal we have to consider how to measure the global scene context appropriately. The representation of the global scene is expected to capture the main layout of the scene. Holistic representations of scenes, termed as the scene-based representations, have been proved to be an efficient method for extracting coarse layout information [24]. Global representation methods are more robust than local appearance representations, because they bypasses some challenging problems, such as local feature detection, feature grouping, and object recognition, which are necessary processing steps in object-based representations. Performance of these intermediate steps has a direct impact on the robustness of the local feature based matching methods. Consequently, we adopt a global scene-based representation method to describe a scene image.

There are three key contributions in this work. First, we argue that traditional template matching is instance-level detection and cannot account for large intra-class divergence, and therefore we propose to use global scene context to improve simple template

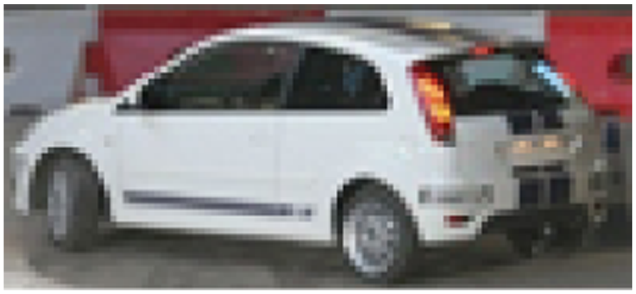


Figure 4. The query image used in the experiments on the Pascal VOC 2008. The size of the query image is 100×46 pixels. The query is cropped from a different image in the Pascal VOC database.
doi:10.1371/journal.pone.0098447.g004

Table 1. Evaluated items of the four datasets.

Dataset	Evaluating items
Pascal VOC 2008	robustness to occlusions, variety of poses, intra-class variations, and background clutter
USC sequence	robustness of occlusions and variety of poses
LabelMe & UIUC	detection performance for both 100% and 25% resolution, generalization
Two videos	non-rigid transform, occlusions, low-resolution

doi:10.1371/journal.pone.0098447.t001

matching. Second, inspired by biological findings, in particular, scene-centered viewpoint and change blindness phenomenon, we propose to measure global scene context using consistency of spatial layout, which is described by the scene-based representations in this work. Third, we propose a scheme to detect a category of objects using a single instance as query by replacing the sliding windows with the query and measuring the change in the global context. The smaller the change in global context is, the more likely there is a similar object at the replaced location. To evaluate the robustness of our method to large intra-class variations, occlusions, a minor variety of poses, low-resolution conditions, background clutter, we perform a detection/localization task on the Pascal VOC 2008 dataset [25], USC sequence dataset [26], LabelMe dataset [27], UIUC car dataset [28], and two videos. The promising results demonstrate the effectiveness of the proposed method.

Related Works. This paper presents a biologically-inspired training-free object detection method using scene context. Some related research are introduced in this section. First, we present conventional template matching methods. Second, we introduce scene-based representations. Finally, we present some approaches that use context to improve computer vision methods

Template matching methods. Conventional template matching has two disadvantages: sensitivity to rotation and computational cost. Many improvements to template matching therefore mainly focus on these two aspects: Ullah and Kaneko [29] have presented a rotation-invariant template matching approach using gradient information in the form of orientation codes as the feature; Yoshimura and Kanade [30] have presented a fast computation method of the normalized correlation for multiple rotated templates by using multi-resolution eigenimages; Choi and Kim [31] have proposed a rotation invariant template matching algorithm based on the combination of the projection method and Zernik moments; Vanderbrug and Rosenfeld [32] have reduced computationally cost by initially using only a sub-template, and applying the rest of the template only when the sub-template's degree of match exceeds a threshold; Krattenthaler et al. [33] have proposed a reduced-cost correlation template matching approach where matching is not performed with the entire template but with a pre-computed set of points of the template.

All the enhancements of template matching are very effective. However, to go beyond instance level detection, i.e., to use a single template to find a broader category of objects, these methods are still too sensitive to large intra-class variations, occlusion, poses and

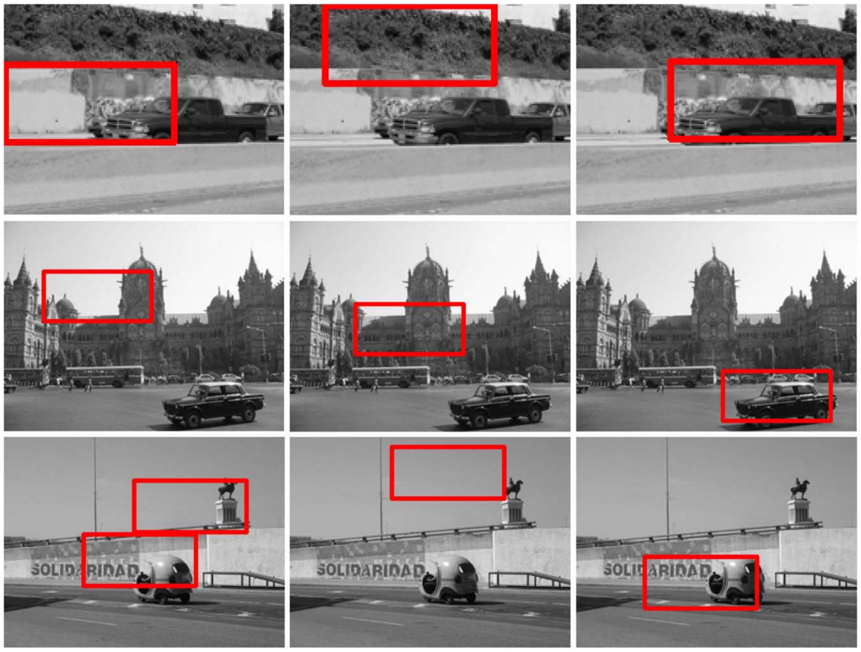


Figure 5. The car detection results with large intra-class variance on the Pascal VOC 2008. For each row, the left image is the result of template matching; the middle image is the results of local *Gist* based template matching; and the right image is the results of the proposed method.
doi:10.1371/journal.pone.0098447.g005

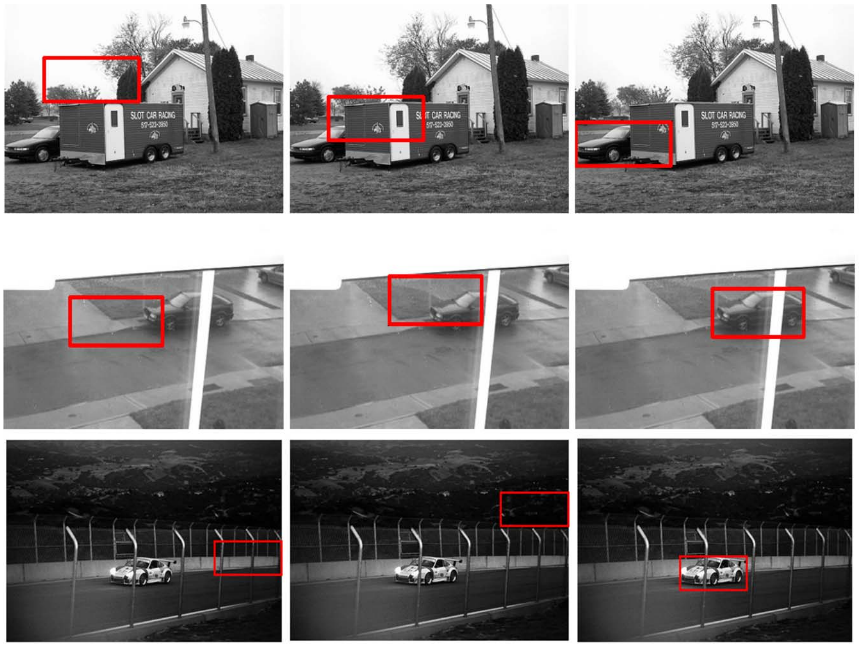


Figure 6. The car detection results with occlusion on the Pascal VOC 2008. For each row, the left image is the result of template matching; the middle image is the results of local *Gist* based template matching; and the right image is the results of the proposed method.
doi:10.1371/journal.pone.0098447.g006

low-resolution conditions, because of the limited robustness of the local appearance features that form the template.

Scene-based representations. Scene-based representations consider scene-images as a whole and describe them as low-dimensional features. Such representations are robust, because (1) they bypass some challenging problems, such as segmentation, feature grouping, and object recognition, which are necessary processing steps in object-based representations; and (2) they average out the local noise by processing globally [24]. The challenge to discover a compact and holistic representation for unconstrained images has prompted significant recent research. The semantic category of most real world scenes can be inferred from their spatial layouts (e.g. a spatial arrangement of the objects and the scene ground plane [16,17]; an arrangement of basic geometrical forms such as simple Geon clusters [6]; or the spatial relationships between regions or blobs of particular size and aspect

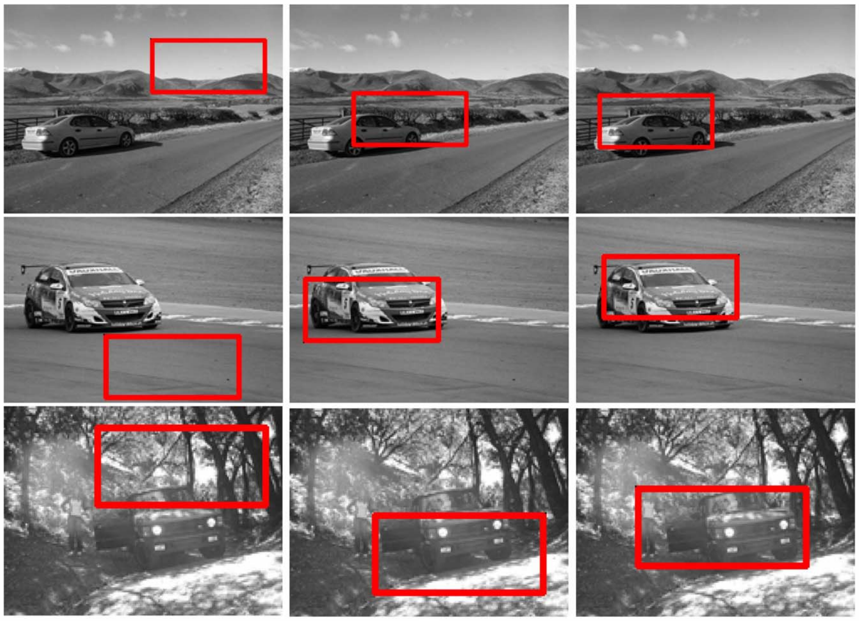


Figure 7. The car detection results with variety of poses on the Pascal VOC 2008. For each row, the left image is the result of template matching; the middle image is the results of local *Gist* based template matching; and the right image is the results of the proposed method.
doi:10.1371/journal.pone.0098447.g007

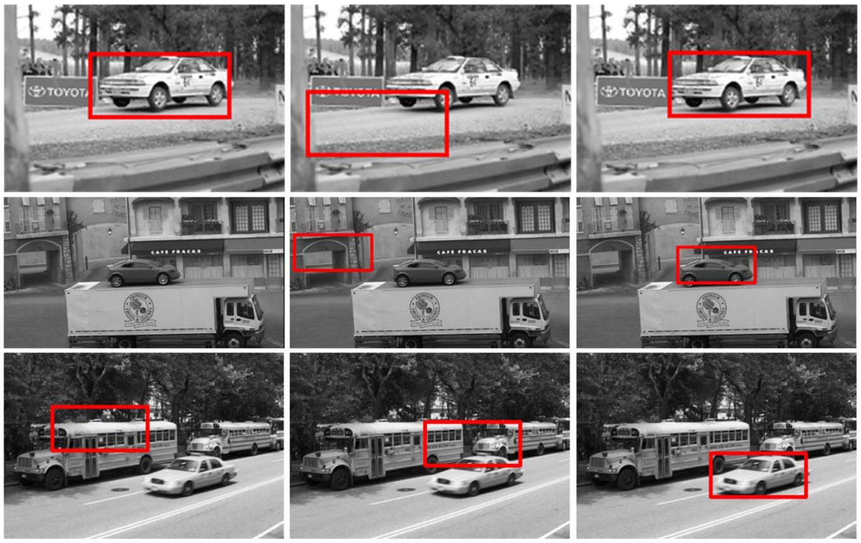
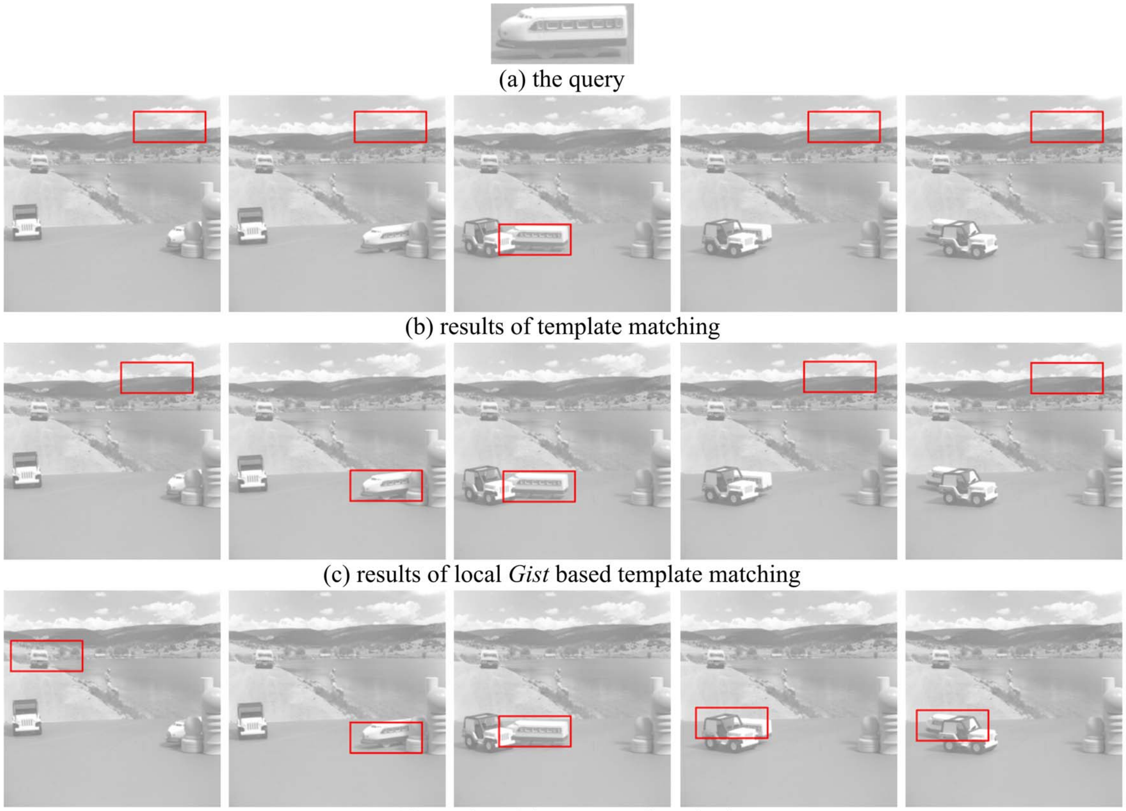


Figure 8. The car detection results with background clutter on the Pascal VOC 2008. For each row, the left image is the result of template matching; the middle image is the results of local *Gist* based template matching; and the right image is the results of the proposed method.
doi:10.1371/journal.pone.0098447.g008

ratio [11,12,17]). Torralba et al. [34] encode some spatial information using steerable wavelet pyramids within individual image sub-regions on a regularly-spaced grid; whilst Siagian and Itti [24] use a multi-scale set of early-visual features instead of steerable wavelet pyramids. Oliva and Torralba [18] introduced a representation for describing the structure of real world scenes,



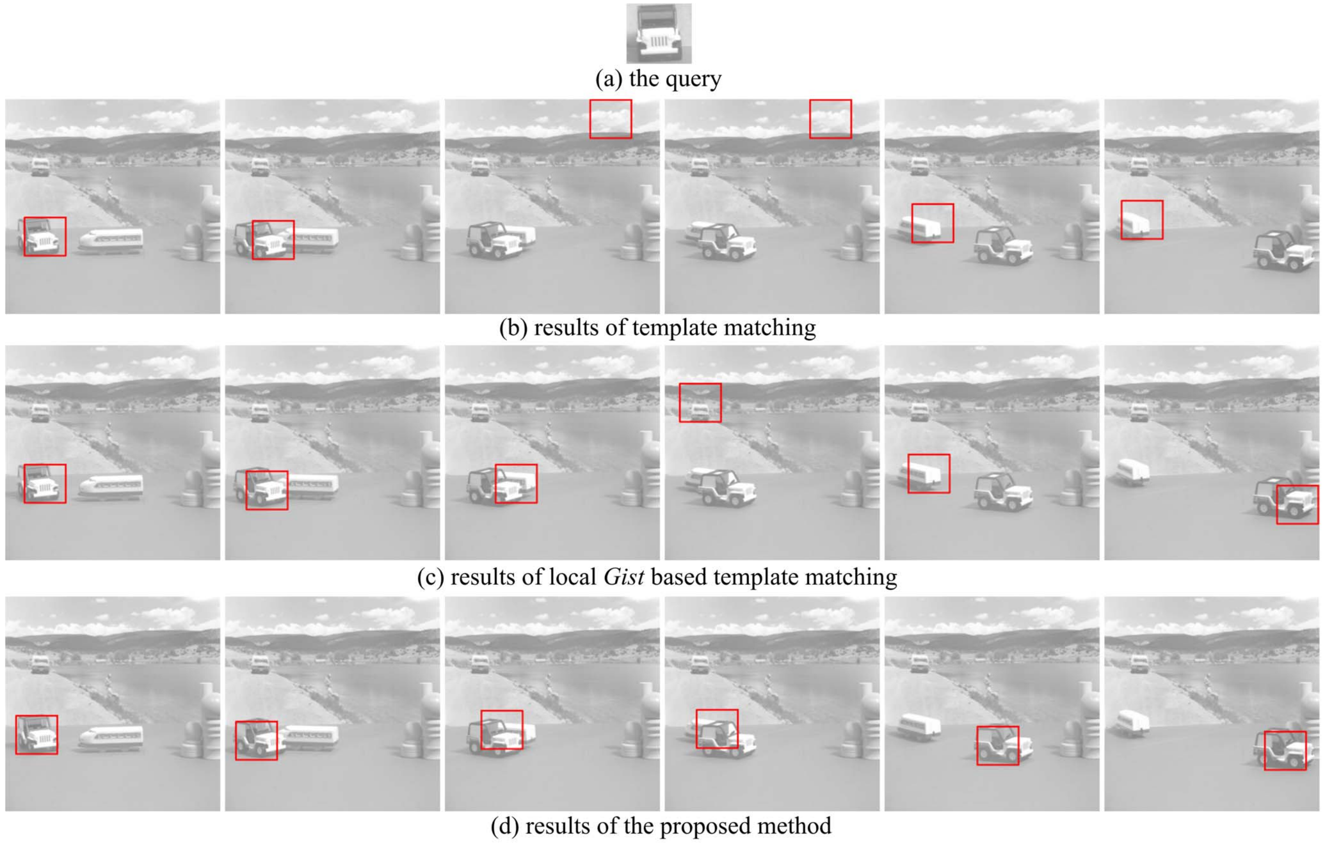
(a) the query

(b) results of template matching

(c) results of local *Gist* based template matching

(d) results of the proposed method

Figure 9. The object detection results of Group 1 on the USC sequence, using the three methods: template matching, local *Gist* based template matching and the proposed method. Five frames of test results are presented, which are Frame 2, 3, 6, 7, 8 of original sequence. (a) shows the query image in the experiments, which is cropped from frame 5; (b) shows the results of template matching; (c) shows the results of local *gist* based template matching; (d) shows the results of the proposed method. For (b), (c) and (d), each image is the detection result for a specific frame.
doi:10.1371/journal.pone.0098447.g009



(a) the query

(b) results of template matching

(c) results of local *Gist* based template matching

(d) results of the proposed method

Figure 10. The object detection results of Group 2 using the three methods: template matching, local *Gist* based template matching and the proposed method. Six frames testing results are presented, which are Frame 5, 6, 7, 8, 9, 10 of original sequence. (a) shows the query image in the experiments, which is cropped from frame 5; (b) shows the results of template matching; (c) shows the results of local *Gist* based template matching; (d) shows the results of the proposed method. For (b), (c) and (d), each image is the detection result corresponding to a frame.
doi:10.1371/journal.pone.0098447.g010

called Spatial Envelope, which is described as a set of perceptual properties (naturalness, openness, roughness, ruggedness and expansion).

In our scene context based detection method, how to represent the scene context appropriately is very important. Scene-centered representations, which are more robust than local appearance representations, may help in this regard. "*Gist*" [34], which extracts coarse layout information of a scene, is used to represent scene images in this work.

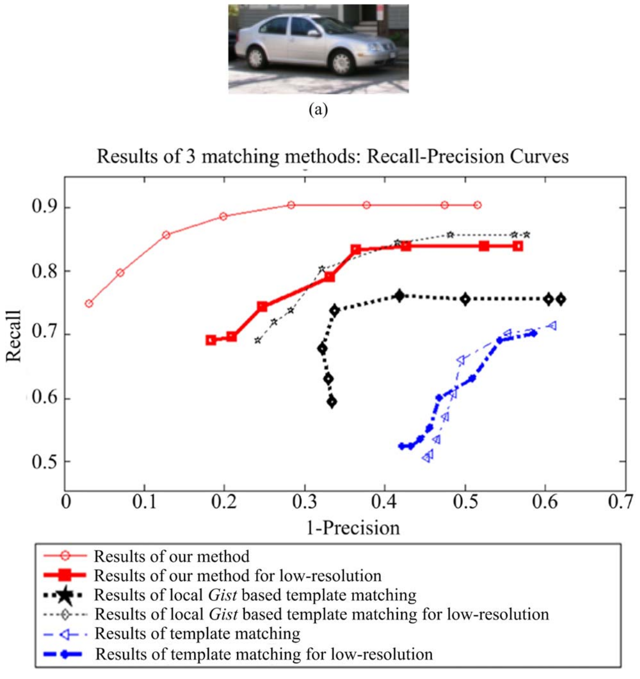
Context-based methods. The relationship between environments and the objects within these environments is very strong in real-world scenes. In typical visual-search experiments, the context of a target is a great hindrance to the detection process [35]. However, the context provides rich information that can serve to help the recognition and detection process [35]. Context-based approaches to object recognition have become increasingly popular over the years. There are three key context representations at different levels: neighbor-based context, object-based context and scene context. Note that, the level of context we used in this paper is scene context. And Colleguilas and Belangie [36] review a variety of approaches for context-based object categorization.

Many context-based methods have demonstrated that context information can be used to improve aspects of computer vision, such as image segmentation, visual attention [37,38], object detection [39], object recognition [40,41], visual tracking [42,43], and scene understanding [38,44,45]. Thus, it is reasonable to apply context to improve template matching methods.

Table 2. The Number of correct detected frames with the three methods for the two groups of experiments on the USC sequence.

Methods	Group 1	Group 2	Missed frames
Template matching	1	2	8
Local *Gist* based template matching	2	4	5
The proposed method	4	6	1

For Group 1 experiments, the results of 5 frames are presented and 6 frames for Group 2. The numbers of missed frames of the three methods are also given.
doi:10.1371/journal.pone.0098447.t002



(a)

(b)

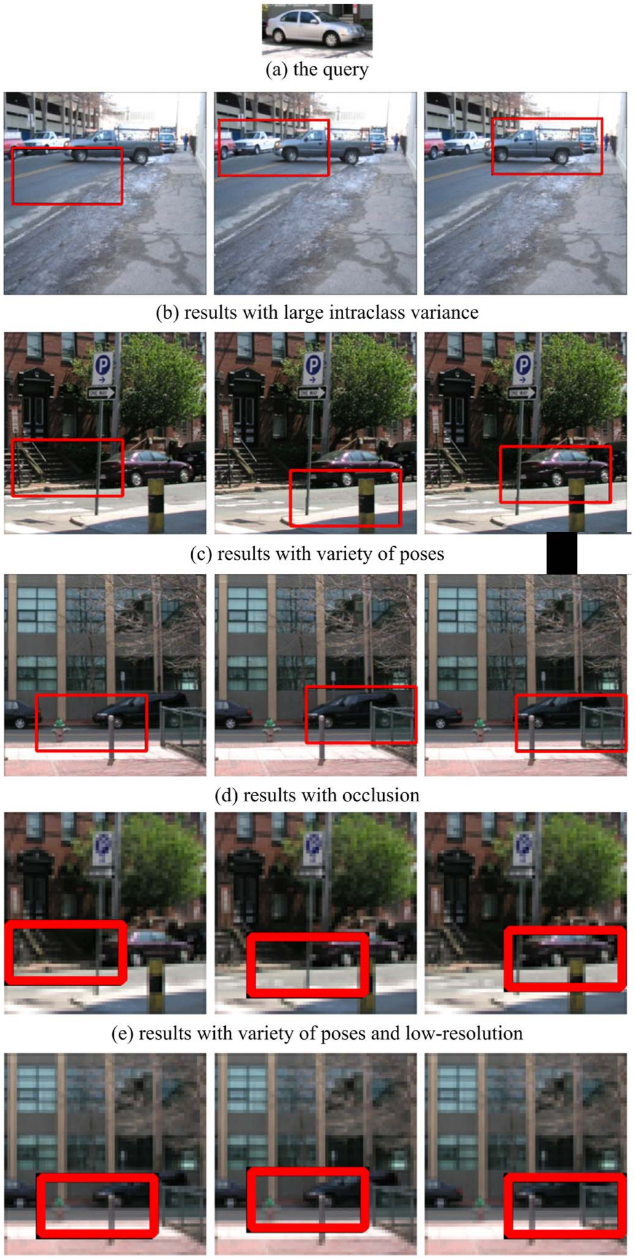
Figure 11. Recall-precision curves of template matching, local *Gist* based template matching, and our method on the LabelMe database. The results for low-resolution images are also given. The results show that our approach outperforms others.
doi:10.1371/journal.pone.0098447.g011

Methods

To locate all instances of an object of interest in a scene image, we slide windows across all positions in the image, as do most object detection approaches. Instead of measuring the similarity of the query and the sub-image in each window directly, we replace the sub-image with the query and measure the change in the global context, as shown in Fig. 2. Then we measure the confidence in how much the query image matches a sub-image using the similarity between the global scene context of the original scene and the replaced scene. Fig. 3 shows the major stages of the proposed method. The main difference between the proposed method and conventional template matching is that traditionally only local information is used for matching, while the proposed method uses global scene context for matching. Due to the ability of global scene context to describe the coarse layout information, the proposed method, to a certain extent, can be considered as a global layout guided matching method.

In this work, *Gist* [18] is applied as the global scene representation. An input image I is filtered using a set of Gabor filters with O orientations and S scales. For each Gabor filter $G_{o,s}$, ($o=1,2,\dots,O$, $s=1,2,\dots,S$), the filtered map can be computed by $M_{o,s}=I*G_{o,s}$. *Gist* cumulates information in an $N\times N$ grid of sub-regions over the maps. Each element of the *Gist* vector is computed by taking the sum over a corresponding sub-region (specified by indices k and l in the horizontal and vertical direction respectively, $k=1,2,\dots,N$, $l=1,2,\dots,N$):

$$v_{o,s,k,l}(I)=\frac{1}{NNWH}\sum_{p=\frac{kW}{N}}^{\frac{(k+1)W}{N}-1}\sum_{q=\frac{lH}{N}}^{\frac{(l+1)H}{N}-1}M_{o,s}(p,q) \quad (1)$$



(a) the query

(b) results with large intraclass variance

(c) results with variety of poses

(d) results with occlusion

(e) results with variety of poses and low-resolution

(f) results with occlusion and low-resolution

Figure 12. Car detection results with different challenges from the LabelMe dataset. For the subfigures (b)–(f), the left images are the result of template matching; the middle images are the results of local *Gist* based template matching; and the right images are the results of our method.
doi:10.1371/journal.pone.0098447.g012

where W and H are the width and height of the entire image, (p,q) is position index. The final $O\times S\times N\times N$ *Gist* vector is composed of all the elements: $V=v_{o,s,k,l}(I)$.

Results and Discussion

Experimental Design

We experimentally compare our method against a conventional template matching method, by performing training-free car

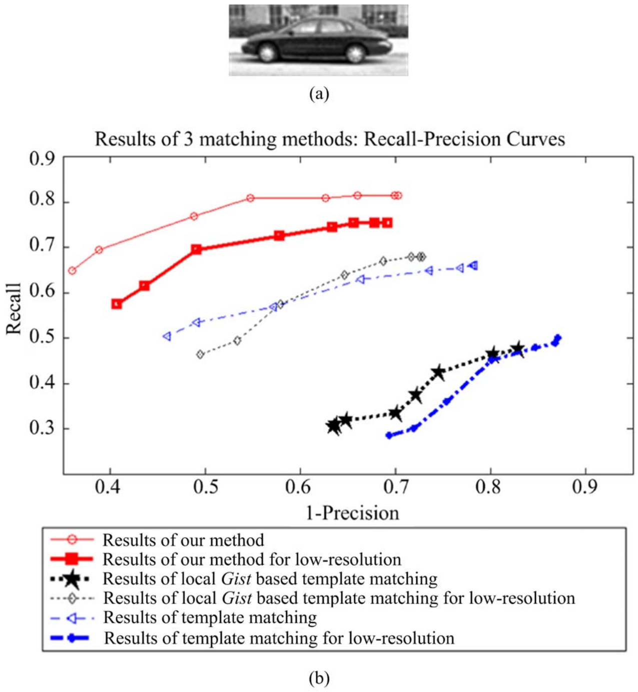


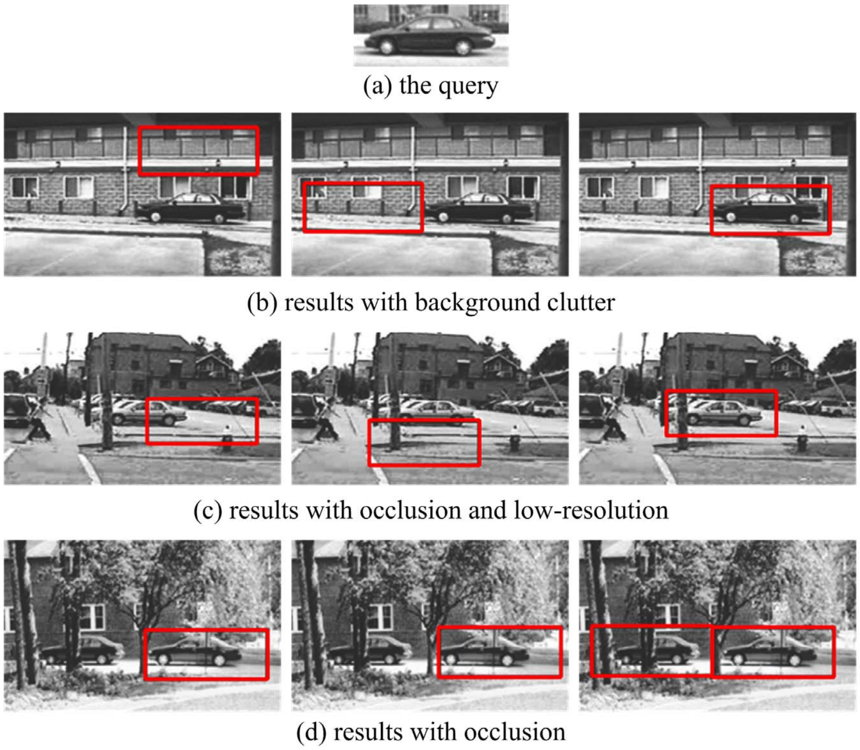
Figure 13. Recall-precision curves for template matching, local *Gist* based template matching and our method on the UIUC car database. The results show that our approach outperforms others.
doi:10.1371/journal.pone.0098447.g013

detection tasks on four well-known datasets: Pascal VOC sets of 2008 [25], USC sequence (one sequence of volume four in USC-SIPI image dataset [26]), LabelMe [27], UIUC [28]. In addition, we test our method on two videos to evaluate a detection-based tracking task. Objects in these datasets can be highly occluded, occupy various poses, include various types within the same category, be low-resolution conditions or have cluttered background. The robustness of detection performance and the generalization of our method are evaluated for items of each dataset (listed in Table 1). Finally, we evaluate the effect of different templates for our method. To sum up, we perform 6 groups of experiments: Experiments on PASCAL, Experiments on USC sequence, Experiments on LabelMe, Experiments on UIUC, Experiments on two videos sequences, and Evaluation of generalization ability (the effect of the different query templates).

Three methods are compared in our experiment on these datasets: (1) template matching method, correlation coefficient of gray values of query and sub-image in a sliding window is used to measure their similarity; (2) local *Gist* based template matching, correlation coefficient of *Gist* of query and sub-image in a sliding window is used to measure their similarity; (3) the proposed method.

Gist descriptor is applied as the global scene representation. In our experiments, a bank of Gabor filters are constructed using 8 orientations and 4 scales, in a total of $4 \times 8 = 32$ filters. The image is divided into a 4×4 grid, and the average energy of each channel in each gird cell is computed. Then the resulting image representation is a $32 \times 4 \times 4 = 512$ dimensional feature vector.

In our experiments, the sliding sub-window moves in steps of 5 pixels horizontally and 2 vertically on all these databases. After



(a) the query

(b) results with background clutter

(c) results with occlusion and low-resolution

(d) results with occlusion

Figure 14. Car detection results with different challenges on the UIUC dataset. For the subfigures (b),(c), and (d), the left images are the result of template matching; the middle images are the results of local *Gist* based template matching; and the right images are the results of the proposed method.
doi:10.1371/journal.pone.0098447.g014

Table 3. The RPC-Equal Error Rates of our approach on the UIUC dataset and the comparison with template matching method, local *Gist* based template matching method, local HoG based template matching method, and the approach proposed by Agarwal et al. [28].

Methods	RPC-Equal error rate (%)
Template matching	60.4
Local *Gist* based template matching	59.1
The proposed method	**77.6**
Local HoG based template matching	62.7
Agarwal et al. [28]	76.5

The highest result is shown in bold.
doi:10.1371/journal.pone.0098447.t003

thresholding the confidences of all positions, "repeated part elimination" [28] algorithm is used to obtain the final positions.

In addition, we performed all processing in grayscale, even when color images are available. Our approach can be easily expanded to a multiple-scale detection task by searching at each scale. However, to highlight the distinguished performance of our method, we just focus on the single scale situation in this work. Thus, the testing images were rescaled before localization/detection so that the car instances are roughly the same size.

Experiments on PASCAL

In this subsection, we evaluate the robustness of our method to large intra-class variations, occlusions, poses and background clutter on the Pascal VOC 2008 dataset. Note that the challenges in each test image are often intertwined. For example, in a testing image, the car may be quite different from the template in both appearance and pose. We report 3 testing results for each challenge. Fig. 4 presents the template used in the experiments. Fig. 5, Fig. 6, Fig. 7, and Fig. 8 show the matching results with large intra-class variations, occlusions, poses, and background clutter respectively. The results of these challenges demonstrate the robustness and effectiveness of the proposed method.

Experiments on USC sequence

For in-depth analysis of our method, we evaluate the robustness to occlusion and variety of poses on a USC sequence, which contains three cars, consists of 10 frames with 512×512 pixels. The challenge of the first group of experiments is occlusion. Five frames with occlusion are used for testing. The second group of experiments investigates the effect of poses. Six frames with a wide variety of poses are tested. The results of these two groups of experiments are shown in Fig. 9 and Fig. 10 respectively. The template matching method, local *Gist* based template matching and our method are compared. The numbers of correctly detected frames are presented in Table 2 for both groups. It can be seen that only one error occurs for our method. The experimental results on the USC sequence show that our proposed method is more robust to occlusion and variety of poses.

Experiments on LabelMe

We used 150 car images to evaluate the performance of the proposed method on the LabelMe database. These images were re-scaled so that the side view of the cars are roughly the same size (138×69). The template is from one of the scene images, as shown

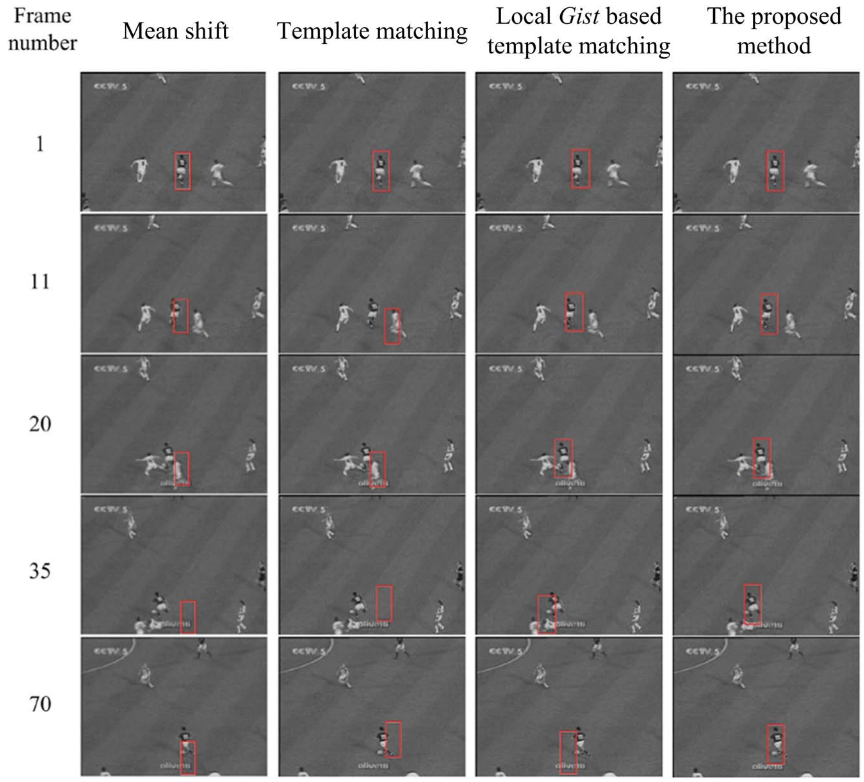


Figure 15. The tracking results for video 1 using the four methods: Mean shift, template matching, local *Gist* based template matching and our method. Five frames are presented, which are Frame 1, 11, 20, 35, 70 of the original video.
doi:10.1371/journal.pone.0098447.g015

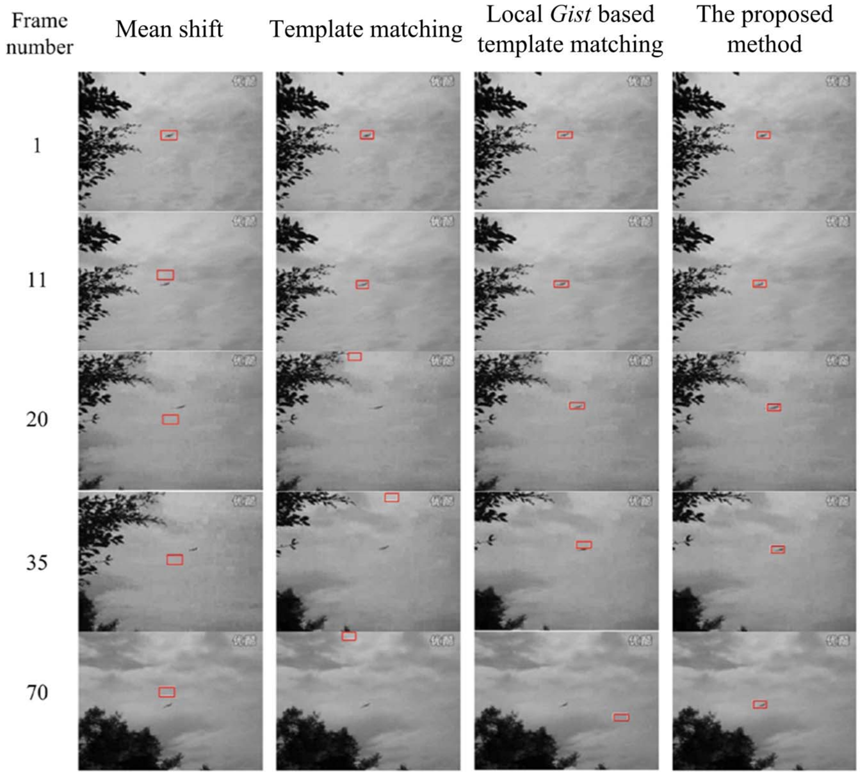


Figure 16. The tracking results for video 2 using the four methods: Mean shift, template matching, local *Gist* based template matching and our method. Five frames are presented, which are Frame 1, 5, 20, 30, 70 of the original video.
doi:10.1371/journal.pone.0098447.g016

in Fig. 11(a). Fig. 11(b) presents the recall-precision curves of template matching, local gist based template matching, and our method, showing that our approach clearly outperforms others. The results for a 25% resolution conditions (the size of the object is 35×18) are given in Fig. 11(b). Some matching results are presented in Fig. 12. These results show the strong robustness of our method to different challenges.

Experiments on UIUC

The UIUC car dataset consists of small (100×40 pixels) training images of 550 cars and 500 backgrounds. There are two sets of test images: a single-scale set in which the cars to be detected are roughly the same size as those in the training images, and a multi-scale set. Our approach can be expanded to multi-scale by searching at each scale, and we focus on the single-scale set in this work. The query template is from the training set on UIUC database, as shown in Fig. 13(a). Fig. 13(b) presents the results of template matching, local *Gist* based template matching and the proposed method as recall-precision curves, showing the good performance of our method. Results from a 25% resolution condition (the size of the object is 25×10) are also given in Fig. 13(b). Some matching results are presented in Fig. 14, which show that our approach is more robust.

We also report the Recall-Precision equal (RPC-equal, i.e., when recall = precision) error rates of the three methods together with that of local HoG [46] based template matching and Agarwal et al. [28] in Table 3. It can be seen that, our method outperforms local HoG based template matching method. The comparison results show that our method can compare with the template matching method with some powerful local representations, such as HoG. And we can see that the RPC-equal error rates of our method and the method in Ref. [28] are about the same. Note that, the latter method is learning based. The results demonstrate that for object detection tasks, the performance of the proposed method using only one object sample is very promising.

Table 4. The number of continually tracked frames (CTF).

Methods	CTF on the first video	CTF on the second video
Mean shift	4	2
Template matching	2	18
Local *Gist* based template matching	28	27
The proposed method	70	70

doi:10.1371/journal.pone.0098447.t004

Experiments on two video sequences

In this subsection, we report the tracking results on two video sequences, to demonstrate the robustness of the proposed method. For each frame, object tracking is very similar to the object detection problem, where the template is the detected result of the previous frame. The first video sequence is from a football match, where the player wearing No. 10 is the target in our tracking task. Tracking is over 70 frames. This task has several difficulties: the target is a non-rigid object; the player is partly occluded during the pass; and the background is moving. Fig. 15 presents the tracking results of template matching, local *Gist* based template matching, and our method. In addition, the results of the Mean shift method are also reported for comparison. Each row of the figure corresponds to a video frame, where the frame number is denoted

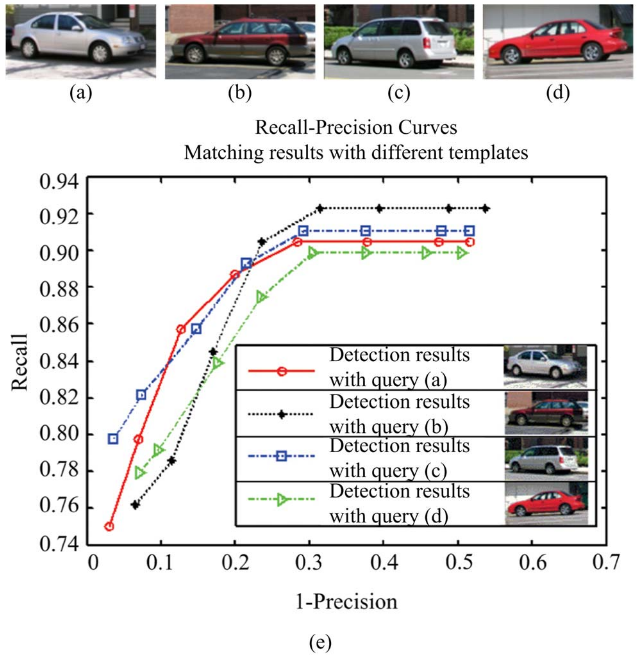


Figure 17. Comparison of performance of our method with 4 different query images on the LabelMe dataset. (a) query 1; (b) query 2; (c) query 3; (d) query 4; (e) Recall-precision curves for our method with 4 different query templates.
doi:10.1371/journal.pone.0098447.g017

at the far left. The tracked locations of the four methods on the frame are shown in different columns.

In the second video sequence, a plane flies in the sky, which is the object of interest. The length of the original sequence was 144 frames. Tracking was initiated in frame 1 and carried out for 70 frames (the plane is in shot for this range of frames). The plane in the video sequence is hard to track with a template matching method, because the plane is very small and hard to identify, and the background is also moving. Fig.16 presents the tracking results of the Mean shift method and the three methods mentioned above. The layout of Fig.16 is identical to that of Fig.15, where each row corresponds to a frame and each column corresponds to a tracking method.

We also report the number of continually tracked frames to evaluate the performance of these methods, as shown in Table 4. These results demonstrate the effectiveness and robustness of our approach for object tracking.

Evaluation of generalization ability

In addition, to analyze the effect of different query images on the performance of our method, we give the recall-precision curves of our method with 4 different query images in Fig. 17 and Fig. 18, on LabelMe and UIUC databases respectively. It can be seen that the detection performance is very close with different query templates. For in-depth analysis, the statistical results (RPC-equal error rates) of our approach with 100 different query templates are reported in Table 5 for both databases. The well-generalized results demonstrate that our approach does not depend on the particularity of queries.

Table 5. Statistical results (RPC-equal error rates) of our method with 100 different query images on the LabelMe dataset and the UIUC car dataset.

Statistical values	LabelMe	UIUC
Mean	91.2	79.5
Standard deviation	0.68	0.34

doi:10.1371/journal.pone.0098447.t005

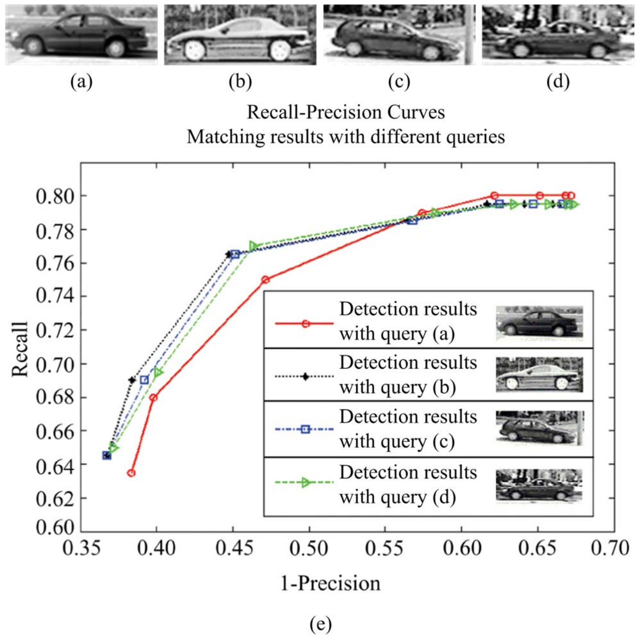


Figure 18. Comparison of performance of our method with 4 different query images on the UIUC car dataset. (a) query 1; (b) query 2; (c) query 3; (d) query 4; (e) Recall-precision curves for our method with 4 different query templates.
doi:10.1371/journal.pone.0098447.g018

Conclusions

In this paper, we have proposed an object detection method using single instances, which measures the similarity of a query and a sub-image in the scene by replacing the sub-image with the query and measuring the change in global scene context based on the *Gist* representation. And we have demonstrated the effectiveness of the proposed method with a variety of experiments. Using scene context, the proposed approach is more robust for object localization, including both detection in static images and tracking in dynamic scenes. Experiments show that our approach outperforms other conventional methods, especially when the queried object may show large intra-class variations, occlusions, various poses, low resolution, or be on a cluttered background.

We believe that the good performance of the proposed method is fostered by the scene context based representation, which describes the coarse layout information of the scene. It suggests that global representation is more robust and an object localization/detection method can be fragile without top-down contextual guidance.

In our current implementation, the method has some limitations: it is sensitive to scale and rotation since we did not take these factors into consideration, which should be relatively easy to incorporate. And our further research will be devoted to removing this limitation.

Author Contributions

Conceived and designed the experiments: CG NS RH. Performed the experiments: CG. Analyzed the data: CG NS RH. Wrote the paper: CG NS RH.

References

1. Shechtman E, Irani M (2007) Space-time behavior-based correlation-or-how to tell if two underlying motion fields are similar without computing them? IEEE Trans. Pattern Analysis and Machine Intelligence 29: 2045–2056.
2. Yeo C, Ahammad P, Ramchandran K, Sastry S (2008) High speed action recognition and localization in compressed domain videos. IEEE Trans. Circuits and Systems for Video Technology 18: 1006–1015.
3. Bar M, Aminoff E, Boshyan J, Fenske M, Gronauo N, et al. (2005) The contribution of context to visual object recognition. Journal of Vision 5: 88a.
4. Biederman I, Mezzanotte RJ, Rabinowitz JC (1982) Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology 14: 143–177.
5. Cox D, Meyers E, Sinha P (2004) Contextually evoked object-specific responses in human visual cortex. Science 304: 115–117.
6. Biederman I (1995) Visual object recognition. In an Invitation to Cognitive Science: Visual Cognition (2nd edition) Kosslyn M and Osherson DN (eds). pp. 121–165.
7. Greene MR, Oliva A (2005) Better to run than hide: Time course of naturalistic scene decisions [abstract]. Journal of Vision 5: 70a.
8. Oliva A (2005) Gist of the scene. In Neurobiology of Attention, L Itti, G Rees and J K Tsotsos (Eds), Elsevier, San Diego, CA.pp. 251–256.
9. Rensink RA, O'Regan JK, Clark JJ (1997) To see or not to see: The need for attention to perceive changes in scenes. Psychological Science 8: 368–373.
10. Simons DJ, Levin DT (1998) Failure to detect changes to people in a real-world interaction. Psychonomic Bulletin & Review 5: 644–649.
11. Oliva A, Schyns PG (2000) Diagnostic colors mediate scene recognition. Cognition Psychology 41: 176–210.
12. Schyns PG, Oliva A From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. Psychology Science 5: 195–200.
13. Henderson JM, Hollingworth A (1999) High level scene perception. Annual Review of Psychology 50: 243–271.
14. Potter MC (1975) Meaning in visual search. Science 187: 965–966.
15. Biederman I (1988) Aspects and extension of a theory of human image understanding. Computational Processes in Human Vision: An Interdisciplinary Perspective, Pylyshyn Z (Ed), Ablex Publishing Corporation: Norwood, New Jersey.
16. Hochberg JE (1968) In the mind's eye. In Contemporary Theory and Research in Visual Perception, RN Haber (Eds), Holt, Rinehart, and Winston: New York.pp. 309–331.
17. Sanocki T, Epstein W (1997) Priming spatial layout of scenes. Psychological Science 8: 374–378.
18. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision 42: 145—175.
19. Levin DT, Simons DJ (1997) Failure to detect changes to attended objects in motion pictures. Psychonomic Bulletin & Review 4: 501–506.
20. Hollingworth A, Henderson JM (2002) Accurate visual memory for previously attended objects in natural scenes. Journal of Experimental Psychology: Human Perception and Performance 28: 113–136.
21. Mitroff SR, Simons DJ, Levin DT (2004) Nothing compares two views: Change blindness can occur despite preserved access to the changed information. Perception & Psychophysics 66: 1268–1281.
22. Simons DJ, Levin DT (1997) Change blindness. Trends in Cognitive Sciences 1: 261–267.
23. Gao C, Sang N (2011) Biologically Inspired Template Matching using Scene Context. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), pp. 1–8.
24. Siagian C, Itti L (2007) Rapid biologically-inspired scene classification using features shared with visual attention. International Journal of Computer Vision 29: 300–312.
25. PASCAL Visual Object Classes Challenge 2008(VOC2008) Results [Online]. Available: http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/index.html.Accessed 2014 May 6.
26. USC-SIPI Image Database [Online]. Available: http://sipi.usc.edu/database/.Accessed 2014 May 6.
27. Russell BC, Torralba A, Murphy KP, Freeman WT (2005) LabelMe: A database and web-based tool for image annotation. Technical report, MIT, AI Lab Memo AIM-2005-025.
28. Agarwal S, Awan A, Roth D (2004) Learning to detect objects in images via a sparse, part-based representation. IEEE Trans. Pattern Analysis and Machine Intelligence 26: 1475–1490.
29. Ullah F, Kaneko S (2004) Using orientation codes for rotation-invariant template matching. Pattern Recognition 37: 201–209.
30. Yoshimura S, Kanade T (1994) Fast template matching based on the normalized correlation by using multi-resolution eigenimages. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS) 3: 2086–2093.
31. Choi MS, Kim W (2002) A novel two stage template matching method for rotation and illumination invariance. Pattern Recognition 35: 119–129.
32. Vanderbrug G, Rosenfeld A (1977) Two-stage template matching. IEEE Transactions on Computers 100: 384–393.
33. Krattenthaler W, Mayer K, Zeiller M (1994) Point correlation: a reduced-cost template matching technique. In Proceedings of IEEE International Conference on Image Processing (ICIP) 1: 208–212.
34. Torralba A, Murphy K, Freeman W, Rubin M (2003) Context-based vision system for place and object recognition. In Proceedings of International Conference on Computer Vision (ICCV), pp. 1023–1029.
35. Oliva A, Torralba A (2007) The role of context in object recognition. Trends in Cognitive Sciences 11: 520–527.
36. Calleguillos C, Belongie S (2010) Context based object categorization: a critical survey. Computer Vision and Image Understanding 114: 712–722.
37. Gao C, Sang N, Huang R (2012) Instance-based attention: where could humans look first when searching for an object instance. Optics Letters 37: 76–78.
38. Torralba A (2003) Contextual priming for object detection. International Journal of Computer Vision 52: 169–191.
39. Murphy K, Torralba A, Freeman W (2003) Using the forest to see the trees: A graphical model relating features, objects, and scenes. Advances in neural information processing systems (NIPS) 16: 1499–1506.
40. Carbonetto P, de Freitas N, Barnard K (2004) A statistical model for general contextual object recognition. In European Conference on Computer Vision (ECCV), pp. 350–362.
41. Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, Belongie S (2007) Objects in context. In Proceedings of International Conference on Computer Vision (ICCV), pp.1–8.
42. Dinh T, Vo N, Medionil G (2011) Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1177–1184.
43. Wu Y, Fan J (2009) Contextual flow. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 33–40.
44. He X, Zemel R, Carreira-Perpinan M (2004) Multiscale conditional random fields for image labeling. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 695–702.
45. Shotton J, Winn J, Rother C, Criminisi A (2007) Texton-boost for image understanding: Multi- class object recognition and segmentation by jointly modeling appearance, shape and context. International Journal of Computer Vision 81: 2–23.
46. Dalal N, Triggs B (2005) Histograms of Oriented Gradients for Human Detection. In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886–893.