SCIENCE CHINA
Information Sciences

• RESEARCH PAPERS •
Special Focus

December 2011 Vol. 54 No. 12: 2461–2470
doi: 10.1007/s11432-011-4487-1

Salient region detection and segmentation for general object recognition and image understanding

HUANG TieJun[1], TIAN YongHong[1]*, LI Jia[2] & YU HaoNan[1]

[1]*National Engineering Laboratory for Video Technology, School of Electrical Engineering and Computer Science, Peking University, Beijing* 100871*, China;*
[2]*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190*, China*

Received June 15, 2011; accepted September 27, 2011

Abstract General object recognition and image understanding is recognized as a dramatic goal for computer vision and multimedia retrieval. In spite of the great efforts devoted in the last two decades, it still remains an open problem. In this paper, we propose a selective attention-driven model for general image understanding, named GORIUM (general object recognition and image understanding model). The key idea of our model is to discover recurring visual objects by selective attention modeling and pairwise local invariant features matching on a large image set in an unsupervised manner. Towards this end, it can be formulated as a four-layer bottom-up model, i.e., salient region detection, object segmentation, automatic object discovering and visual dictionary construction. By exploiting multi-task learning methods to model visual saliency simultaneously with the bottom-up and top-down factors, the lowest layer can effectively detect salient objects in an image. The second layer exploits a simple yet effective learning approach to generate two complementary maps from several raw saliency maps, which then can be utilized to segment the salient objects precisely from a complex scene. For the third layer, we have also implemented an unsupervised approach to automatically discover general objects from large image set by pairwise matching with local invariant features. Afterwards, visual dictionary construction can be implemented by using many state-of-the-art algorithms and tools available nowadays.

Keywords object recognition, image understanding, visual saliency, salient object segmentation, visual dictionary

Citation Huang T J, Tian Y H, Li J, et al. Salient region detection and segmentation for general object recognition and image understanding. Sci China Inf Sci, 2011, 54: 2461–2470, doi: 10.1007/s11432-011-4487-1

1 Introduction

General object recognition and image understanding is widely recognized as a very difficult problem in computer vision and multimedia retrieval. In computer vision community, there are an abundance of techniques and systems for solving various well-defined vision tasks. However, most of them are task-specific and seldom can be generalized to a wide range of applications. While in multimedia retrieval community, it was already observed that the general media understanding and retrieval problem was bottlenecked by the semantic gap [1]. Over the past few years, the whole community has been working very hard to address this issue but even the most cutting-edge image understanding system can only

*Corresponding author (email: yhtian@pku.edu.cn)

interpret an image with limited vocabularies/concepts (hundreds or less) and without a satisfactory precision. It is thus expected that there shall be some significant break-through to be made in developing effective solutions to this general and broad problem.

Local invariant features such as SIFT [2] and SURF [3] are one of the most prominent tools for object recognition in the last decade. The local feature is invariant to affine transforms and possible photography distortions. The object recognition problem is then converted to match local features extracted from a given image with that of the target object under geometric constraint. The underlying rationalities are 1) to identify an object and distinguish it from others, local features are better than other features in terms of object representation; 2) the elaborately-designed local features can keep invariant under various transforms and distortions which exist inevitably in object recognition and image understanding; 3) the geometric relationship among these local features is also an important feature to represent and identify an object. Moreover, a group of local features with a certain geometric relationship can represent an object uniquely, consequently leading to a feasible solution for the general object recognition problem.

The local invariant feature has been utilized for image understanding in recent years as well. By clustering a large number of local features extracted from a quantity of training images, the cluster centroids are taken as visual words which are used to represent an image as bag-of-visual-words (BoWs). Then different models can be employed to map the visual-words with a set of predefined visual concepts, which are in turn used to index images. This is true for natural language in which a "word" is the fundamental block for all sentences and then concepts. However, for the natural image, the assumption that the high-level visual concepts share a common set of visual words is questionable. Moreover, the semantic gap still cannot be bridged by employing the BoW representation derived from local invariant features. This is because the BoW representation loses the most important information that can be used to identify an object when clustering local features to visual words. In our opinion, a break-through approach for image understanding must face squarely the visual objects (including rigid objects such as buildings and soft objects such as different kinds of clouds). The meaningful elements for high-level concepts and semantic understanding are the visual objects rather than the low-level features or its simple derivatives (e.g., BoWs).

In this paper, we propose a selective attention-driven model for general image understanding, named GORIUM (general object recognition and image understanding model). The key idea of GORIUM is to discover recurring visual objects in multiple images by selective attention modeling and pairwise local invariant features matching, and then construct a visual dictionary (VDic) to interpret any new image. Unlike most of existing systems which compare a given image with specific object(s), GORIUM does not specify any object in advance but discover the object occurrences in a given image set.

The extraction of interesting objects is the fundamental task in GORIUM. Since visual saliency can serve as one sort of selection mechanisms to pop out important contents, it is possible to exploit visual saliency for interesting object detection and segmentation. By effectively utilizing visual saliency for detection and segmentation of interesting objects, GORIUM can be formulated as a four-layer bottom-up model, i.e., salient region detection, object segmentation, automatic object discovering and visual dictionary construction. The lowest layer exploits visual saliency to detect salient objects in an image. In this work, we have proposed two learning frameworks for visual saliency estimation, namely multi-task learning and rank learning. Given the estimated visual saliency, the second layer exploits a simple but effective learning approach to generate a sketch map and an envelop map from several raw saliency maps. As such, the most confident parts of saliency maps can be utilized to segment the salient objects precisely from a complex scene. For the third layer, we have also implemented an unsupervised approach to automatically discover general objects from large image set by pairwise matching local invariant features of the images. Afterwards, visual dictionary construction can be implemented by using many state-of-the-art algorithms and tools available nowadays.

The remainder of this paper is organized as follows: Section 2 presents the framework of the GORIUM model. We summarize our recent work on learning-based visual saliency estimation in section 3 and saliency-driven object segmentation in section 4. Then the possible solution for implementing automatic object discovering and visual dictionary construction will be described in section 5. Section 6 will conclude

this paper and discuss the future work.

2 The GORIUM model

As a start point, a large set of images in a given domain should be collected and fed into the GORIUM model to train the VDic. Currently, such an image set can be easily obtained from the Web (e.g., ImageNet [4]). By performing pairwise feature matching operations on the image set, GORIUM could discover various visual objects which appear in multiple images repeatedly. A basic question is if we expect GORIUM to effectively recognize and understand all kinds of photos captured from our natural world, how many visual objects it should discover and collect into its VDic. Irving Biederman evaluated that about 30000 different visual objects are recognized in the course of one person's lifetime [5]. Therefore, a reasonable goal for GORIUM is to recognize and collect 100000 or more visual objects-three times or more than a person can do. Considering billions of photos are available online and there are millions of visual objects inside, to discover more than 100000 frequent objects is not an over-ambitious goal.

In general, the GORIUM model performs four steps in a bottom-up manner to discover interesting objects and then construct the VDic:

1) Detect salient region. To find the most salient area of a given image, the underlying assumption is that an object in that image is more visually salient than the background. Moreover, all potential objects appearing in the VDic should be present in some images as the most salient region. Considering there are millions of images in the training set, the assumption makes sense. In other words, even some objects are missed if they are not the most salient region in an image, the following object discovery step still can find them from their salient instances in other images. Note that here we use the term "region" to denote the salient components of an image which roughly corresponds to an object or part of an object. In our design, GPOIUM holds no priors about "object" in this step. The only thing it can do is to understand which region of the image is more salient.

2) Segment object region from background. Segmentation is to pop potential objects out from the background. This task is made difficult by the wide variability of the object's shape, appearance, and its surrounding complex scene. Typically, values on a saliency map can serve as beliefs of pixels' labels and thus are highly useful for segmentation. Such saliency priors can be exploited by various strategies to segment objects from visual scenes. This paper will briefly introduce our saliency-based segmentation method which shows high accuracy in two benchmark datasets. However, the accuracy of cutting a potential visual objects from an image is not the precondition for discovering objects from millions of images. In other words, the coming pairwise-objects matching is not so sensitive to slightly inaccurate segmentation results.

3) Discover façades of objects from images. As the projection of a 3D scene to a 2D plane, an image only retains the appearance of an object from one view. To distinguish from the term "object" itself, this paper refers to the projection of an object in an image as "façade". Literally, façade is used to express different sides of a building in several images captured at multiple orientations. Here we extend its meaning to express the appearance of any 3D object in a 2D image. We can perform pairwise salient object/region matching to discover the recurring façade candidates, and then several visually similar façade candidates can be grouped and merged into one façade. Finally, only high frequent façades appeared in multiple images are chosen as the entities in the VDic.

4) Define objects for the VDic. In the final step, the discovered façades are used to define entries (i.e., visual objects) of the VDic. In general, an isolated façade which rarely shares local features (e.g., keypoints) with others will define a planar object. Two typical examples are outdoor logos and traffic signs. Meanwhile, if two or more façades overlap side by side (which could be found if two façades share multiple keypoints in one side area), they may be different façades of one object. Such a typical example is landmark building. The visual attributes of the façade(s) could be colligated as the visual description of the defined object. In this way, a visual object in the dictionary can be defined by the façade(s) extracted from multiple images and characterized by the keypoint set with spatial relationship and the appearance attributes inheriting from the corresponding façade(s). Moreover, we can also learn the textual semantic

concepts of the object if some of the original images are provided with textual annotations.

Therefore, the GORIUM model can be expressed as a selective attention-driven general image understanding model with four layers, namely salient region detection, object segmentation, automatic object discovering and visual dictionary construction. Each layer roughly corresponds to one step in the VDic construction process described above. Note that the VDic construction is an incremental and dynamic process. That is, if there are new images available, the VDic can be easily upgraded by incremental learning.

Once generated, the VDic can be used to analyze and interpret a query image by the following steps: 1) extract local features from the query image; 2) query the VDic to find the vocabulary entity by matching the local features under geometric constraint; 3) (optional) compare the appearance attribute(s) of the query image and the entities in the VDic; 4) interpret the query image using the matched visual objects-their positions and spatial relationships, their textual description and appearance attributes.

3 Learning-based salient region detection

Typically, the visual world is highly structured in a 3D manner. As a result, the scenes in images are composed of features that are not random in the 2D plane. Neurobiological evidence shows that the stable properties of the visual environment, such as rough spatial layout and predicable variations, can work as the contextual priors for individuals to find the target in similar environments. In neurobiology, this phenomenon is called contextual cueing effect [6]. This effect can be represented as adopting similar task-related "stimulus-saliency" functions and model fusion strategies in similar scenes. In our work, these functions and fusion strategies are learnt from scenes in the training data and then transferred to new scenes to estimate their visual saliency maps. Towards this end, we have proposed two learning frameworks for visual saliency estimation, namely multi-task learning and rank learning.

3.1 Multi-task learning for visual saliency estimation

In [7], we propose a probabilistic multi-task learning framework for visual saliency estimation. To the best of our knowledge, it is the first approach that explores the problem of visual saliency computation with the multi-task learning framework. This framework can adaptively select different fusion strategies for different scenes to integrate the predictions made by the bottom-up and top-down models.

As shown in Figure 1, this framework mainly consists of two modules: the bottom-up module and the top-down module. In the bottom-up module, the low-level processes of human vision system are simulated through multi-scale wavelet decomposition and mutual feature competition. In contrast, the top-down module adopts a learning-based approach to simulate the influence of high-level processes in human vision system, which may bias the mutual competition between input visual signals. In the learning process, a multi-task learning algorithm is proposed to simultaneously learn the task-related "stimulus-saliency" mapping functions and fusion strategies for different scenes. Given K training scenes each with N blocks, we have $K \times N$ input-output pairs $(\boldsymbol{x}_{kn}, y_{kn})$. Then the optimization problem can be formulated as

$$\begin{aligned} &\min_{\mathbb{F},\mathbb{W}} \frac{1}{KN} \sum_{k=1}^{K} \sum_{n=1}^{N} l\left(\sum_{t\in\mathbb{T}} c_{kt} \hat{f}_t(\boldsymbol{x}_{kn}) + \alpha_{kb} b_{kn}, y_{kn} \right), \\ &\text{s.t. } 0 \leqslant c_{kt}, \alpha_{kb} \leqslant 1, \quad \sum_{t\in\mathbb{T}} c_{kt} + \alpha_{kb} = 1, \quad \forall\, k \in \{1, \ldots, K\}, \end{aligned} \tag{1}$$

where $c_{kt} = \sum_{i=1}^{T_k} \alpha_{ki}\beta_{kit}$, $\beta_{kit} = p(\hat{\mathcal{T}}_t|\mathcal{T}_{ki})$ is the probability that the complex task $\mathcal{T}_{ki}$ comprises the typical task $\hat{\mathcal{T}}_t$, α_{kb} is the probability that the bottom-up process $\mathcal{T}_{kb}$ controls the visual attention on the kth scene, and b_{kn} is the probability that the nth block (characterized with low-level descriptor $\boldsymbol{x}_{kn}$) is the salient block in such process, $l(\cdot)$ is a predefined loss function which quantifies the cost of predicting saliency for the input $\boldsymbol{x}_{kn}$ with the "ground-truth" y_{kn}, and $\hat{f}_t(\boldsymbol{x}_{kn})$ is a task-related "stimulus-saliency" mapping function. We can further incorporate several penalty terms to enforce the sharing of information

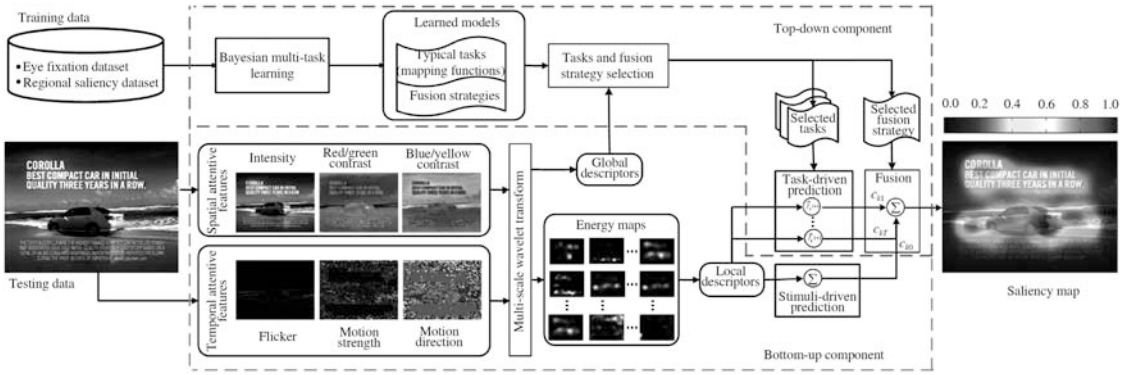

Figure 1 The probabilistic multi-task learning framework for visual saliency estimation.

between tasks to improve the robustness of the learned parameters. Extensive experiments on two public eye-fixation dataset (MTV and ORIG) and one regional saliency dataset (RSD) show that our approach (PMTL) outperforms previously reported work remarkably-seven bottom-up approaches and one top-down approach [7].

3.2 Rank learning for visual saliency estimation

In the contextual cueing effect process, a requisite step is to identify the search priority of each location. Such priority is closely related to visual saliency and can be derived from local visual attributes and pairwise contexts. Therefore, we can formulate visual saliency estimation as a rank learning problem to estimate the searching priority of each location. To learn a "stimulus-saliency" function to distinguish salient targets and distractors on the sparsely labeled eye-fixation dataset, we propose a cost-sensitive rank learning approach to avoid the explicit selection of positive and negative samples [8]. Furthermore, by integrating the multi-task learning and rank learning frameworks, we have also proposed a multi-task rank learning approach to infer multiple saliency models for different scene clusters [9]. Again, they are among the first work to introduce the rank learning framework to solve the visual saliency computation problem.

In our multi-task rank learning approach, scenes with similar contents are first grouped into the same cluster; for each cluster, a ranking function is optimized to give ranks for all subsets in a scene so that these estimated ranks are expected to approximate the ground-truth ranks. Without loss of generality, we define $\phi_m(\boldsymbol{x}) = \omega_m^{\boldsymbol{T}}\boldsymbol{x}$ since various pre-attentive visual features are often integrated into experienced wholes with linear weights for saliency estimation. Let $\boldsymbol{W}$ be an $L \times M$ matrix with the mth column equal to ω_m. Then the optimization objective of multi-task rank learning can be defined as

$$\begin{aligned} &\min_{\boldsymbol{W},\alpha} \mathcal{L}(\boldsymbol{W},\alpha) + \Omega(\boldsymbol{W},\alpha), \\ &\text{s.t.} \sum_{m=1}^{M} \alpha_{km} = 1, \forall k \quad \text{and} \quad \alpha_{km} \in \{0,1\}, \ \forall m, \end{aligned} \tag{2}$$

where $\mathcal{L}(\boldsymbol{W},\alpha)$ is the empirical loss and $\Omega(\boldsymbol{W},\alpha)$ is the penalty term that encodes the prior knowledge on the parameters. The empirical loss $\mathcal{L}(\boldsymbol{W},\alpha)$ can be defined in a pairwise manner:

$$\mathcal{L}(\boldsymbol{W},\alpha) = \sum_{k=1}^{K}\sum_{m=1}^{M} \alpha_{km} \sum_{u\neq v}^{N} [g_{ku} < g_{kv}]_{\boldsymbol{I}} [\omega_m^{\boldsymbol{T}}\boldsymbol{x}_{ku} \geqslant \omega_m^{\boldsymbol{T}}\boldsymbol{x}_{kv}]_{\boldsymbol{I}}, \tag{3}$$

where $[x]_{\boldsymbol{I}} = 1$ if x holds, otherwise $[x]_{\boldsymbol{I}} = 0$. $\Omega(\boldsymbol{W},\alpha)$ can be written as the weighted linear combination of the penalties that encode the priors on scene clustering, model correlation and model complexity [9]. With the appropriately-defined empirical loss and penalty term, we can use the EM algorithm to iteratively solve the problem and ensure the convergence. We compare the proposed approach with the

state-of-the-art saliency models (including seven bottom-up models and three top-down models) and three ranking models on a public eye-fixation dataset. Experimental results show that our multi-task rank learning approach outperforms these methods remarkably in visual saliency estimation [9].

4 Saliency-driven object segmentation

Given the estimated visual saliency, the next problem is how to exploit it to precisely segment objects from complex scenes. In this case, values on a saliency map can serve as beliefs of pixels' labels and thus are highly useful for segmentation. Typically, saliency-based segmentation methods are free from human interaction and thus can be applied to any large dataset. However, visual saliency cannot guarantee too much the accuracy of object segmentation. For example, some saliency-based methods are sensitive to local sudden changes in the background (e.g., [7, 10–12]). In these methods, distracters might be treated as salient objects due to their high saliency values. This causes each segmentation result to be an envelope-like area containing the objects. Meanwhile, there are also some other methods (e.g., [13]) that prefer to highlight only some important parts of objects (referred to as sketch). Although objects segmented by the two kinds of saliency-based methods independently may suffer some problems, it is possible to obtain desirable results by integrating them in a unified framework.

Towards this end, we introduce the concept "complementary saliency map" (CSM) which consists of an envelope map and a sketch map. As shown in Figure 2, the envelope map always highlights a large area containing the objects while the sketch map prefers to highlight small areas inside each salient object. Then pixels with low saliency in the envelope map can be regarded as background seeds while pixels with high saliency in the sketch map can be treated as object seeds. As such, only the most confident parts of CSMs are utilized for object segmentation. This decreases the ambiguity of saliency maps in existing saliency-based segmentation methods. In [14], we employ an ad hoc way to generate sketch and envelope maps; then two KD-trees are built as the object-background color model and the remaining pixels are classified using this model. However, one limitation of the approach is the sketch and envelope maps are specified in an ad hoc way, making it difficult to directly apply them to different datasets, especially those with complex scenes.

To address this problem, we combine different raw maps [10–12, 15–17] in a "mixture-of-experts" framework to generate robust CSMs. By raw maps, we refer to saliency maps that are directly generated from a single saliency model and do not suffer complex post-processing operations. If we consider each raw map as an expert, the saliency value of a pixel in the map will reflect the expert's belief of this pixel being salient; when several experts are mixed together, their opinions can be integrated to generate CSMs. As such, we can use add and multiply operations to simulate the process of integrating raw maps. Let $\mathbb{R} = \{R^{(k)}\}_k$ be a set of K raw maps for image I, with the corresponding envelope weight matrices $\mathbb{W}_{\text{env}} = \{W_{\text{env}}^{(k)}\}_k$ and sketch weight matrices $\mathbb{W}_{\text{ske}} = \{W_{\text{ske}}^{(k)}\}_k$. Then for each pixel p in envelope map S_{env} and sketch map S_{ske}, its value can be determined according to the following models:

$$\begin{aligned} S_{\text{env}}(p) &= \sum_{k=1}^{K} R^{(k)}(p) W_{\text{env}}^{(k)}(p), \\ S_{\text{ske}}(p) &= \prod_{k=1}^{K} (R^{(k)}(p))^{W_{\text{ske}}^{(k)}(p)}. \end{aligned} \tag{4}$$

We observe that the ground-truth map can serve as the ideal envelope map and sketch map simultaneously. Naturally, we want CSMs to fit it as well as possible. This is a multiple linear regression problem. In general, least squares method is a standard approach for approximate solution of over-determined systems. In our case, we try to minimize the sum of squared errors (SSE) for each pixel p:

$$SSE(p) = \sum_{n=1}^{N} \left(\sum_{k=1}^{K} (R^{(n,k)}(p) \cdot W_{\text{env}}^{(k)}(p)) - G^{(n)}(p) \right)^2. \tag{5}$$

Eq. (5) is a convex function and we can easily derive its analytical solution.

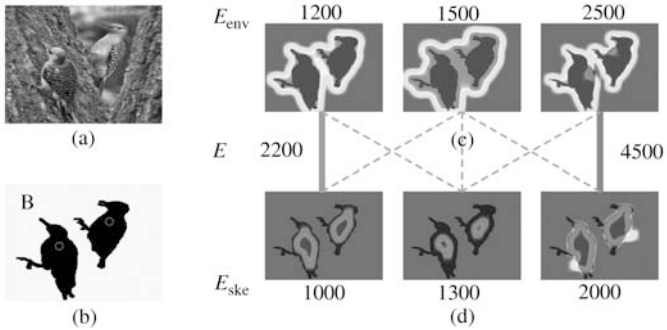

Figure 2 Demonstration of complementary saliency maps. (a) Original image; (b) the ground-truth with "Obj" and "Bkg" marked; (c) some examples of envelope maps; (d) some examples of sketch maps.

Table 1 Performances of various saliency-based methods on the SOSB and MOCB datasets[a)]

Algorithm	SOSB				MOCB			
	Precision	Recall	F-measure	IMP (%)	Precision	Recall	F-measure	IMP (%)
Itti98 [10]	0.78	0.49	0.60	50.0	0.48	0.47	0.47	66.0
Hou07 [12]	0.67	0.57	0.62	45.2	0.50	0.51	0.50	56.0
Achanta09 [11]	0.82	0.75	0.78	15.4	0.49	0.49	0.49	59.2
Harel07 [16]	0.71	0.72	0.71	26.8	0.58	0.60	0.59	32.2
Seo09 [17]	0.66	0.55	0.60	50.0	0.53	0.58	0.55	41.8
Goferman10 [15]	0.58	0.69	0.63	42.9	0.58	0.67	0.62	25.8
Yu10 [14]	0.88	0.89	0.88	2.2	0.74	0.64	0.69	13.0
The proposed	**0.91**	**0.90**	**0.90**		**0.76**	**0.80**	**0.78**	

a) The best performance for each metric is marked as the bold figure.

Then segmentation can be done using the graph cuts technique which provides a powerful framework to combine CSMs, appearance affinity and neighboring interaction into the object segmentation process. In general, it seeks a labeling assignment l that globally minimizes the following cost function (see [18]):

$$C(l) = \sum_{\{p,q\}\in\mathcal{N}} V_{p,q}(l_p, l_q) + \sum_{p\in I} D(l_p), \tag{6}$$

where $\mathcal{N}$ stands for the neighborhood system, $V_{p,q}(l_p, l_q)$ is a binary cost to guarantee that two neighboring pixels p and q are likely to have the same label, and $D(l_p)$ is a unary cost for punishment of assigning a certain label to pixel p. An appearance model used for computing $D(l_p)$ should well represent the visual feature distribution in a certain area. In our work, we set a threshold to binarize the envelope map. Pixels whose saliency values are above (below) the threshold are used to train object (background) GMMs. The optimal solution of the graph cuts can be solved by a fast max-flow algorithm [19].

We conduct comprehensive experiments to examine whether the proposed approach can segment objects in various visual scenes. The two datasets are denoted by SOSB and MOCB, respectively. The SOSB dataset consists of 1000 one-object images with various object classes, which were selected and labeled with exact object masks by Achanta et al. [11]; while the MOCB dataset contains 1474 images selected from [20] and PASCAL VOC09, in each of which various objects coexist in a complex background. The overall performances of various methods are shown in Table 1. On SOSB, our approach has achieved rather good performance, with F-measure of 0.90. While on MOCB, the performances of all the segmentation methods, including ours, decrease largely from simple scenes to complex scenes. In this case, our approach still outperforms all the comparison methods, with F-measure of 0.78. Some representative results on the two datasets are illustrated in Figure 3.

5 Visual dictionary construction

With the segmented objects or regions, GORIUM can perform pairwise salient object/region matching

Figure 3 Representative results on SOSB and MOCB. (a) Original images with ground-truth; (b) [10]; (c) [12]; (d) [11]; (e) [15]; (f) ad hoc envelope maps from [14]; (g) ad hoc sketch maps from [14]; (h) our envelope maps; (i) our sketch maps; (j) our final segmentation results.

to discover the recurring façades, and then the discovered façades are used to define entries (i.e., visual objects or units) of the VDic. In this section, we will discuss the possible solution for implementing automatic façade discovering and visual dictionary construction.

5.1 Façade discovery by pairwise matching

As a general framework for object recognition and image understanding, GORIUM should be able to learn from any training image set in a specific domain. What we are doing is to discover thousands of visual objects from a publicly available dataset—ImageNet, which collects more than 12 million images and about 5% of them are labeled with bounding boxes [20].

The image matching methods used by GORIUM are similar with those for matching a given image with pre-defined visual object(s). The only difference is the salient objects/regions in all images are matched pairwise with each other. The procedure is summarized as follows:

1) Extract local features (e.g., keypoints) in salient regions and create a local features set for each image.

2) Find all matched keypoints in all image pairs, based on the similarity of local features at these regions.

3) Check geometry constraint for the matched keypoints, e.g., using the generalized Hough transform.

4) If there are enough keypoint pairs satisfying geometry constraint, then the two corresponding salient regions in an image pair are chosen as a façade candidate. In this case, the façade candidate can be represented by merging the two salient regions according to the geometry mapping relationship.

5) After a round of pairwise matching, similar façade candidates are grouped into one façade. Only high frequent façades appeared in multiple images can be chosen as the entities in the VDic.

Note that the geometry constraint means that two groups of matched keypoints should keep their geometric isomorphism relationship unchanged under different affine transformations such as resizing, rotation, cropping and aspect-ratio changes. In abstract algebra, such an isomorphism is a kind of mapping between objects that shows a relationship between two properties or operations (e.g., structurally identical).

The main problem here is the complexity of pairwise matching on a large set. Nowadays the matching between two images on a normal PC needs one or more seconds. The pairwise matching of one billion images will spend 10^{18} seconds or 10^{11} years on a PC. Even when using the fastest super-computer nowadays which is about 10^5 times faster than a PC, it still needs one million years. If the number

of images decreases to one million, 100 years are needed for a PC and one year needed for the supercomputer. As a result, a fast algorithm is necessary for the pairwise matching.

In fact, this pairwise matching problem can be cast as an image search task. By employing inverted index technique and tree-based search algorithms, one query in a million-scale image set can be completed in one second on a normal PC server [21]. Our landmark searching system can also perform one query in one second in the one million-scale landmark image database [22]. That is to say, the pairwise matching on one million images could be completed in one million seconds or about 10 days by a PC server. Currently, we are optimizing the search algorithm to deal with 12 million images of ImageNet.

5.2 Visual dictionary construction

In the VDic, each vocabulary entry often corresponds to a visual object that is defined by a façade. If a façade has very few similar keypoints with other discovered façades, we can assume that it corresponds to a planar or tabular object. In this case, an isolated object which equals the façade becomes a new entry of the VDic. Note that the façade may be synthesized by a group of façade candidates and each candidate may be further synthesized by multiple similar salient regions from the input image set.

Composite object is another type of the entries in the dictionary. It comes from multiple façades which could be stitched one by one with a set of shared keypoint pairs. Again, keypoints matching algorithm can be carried out to find the spliced side area between any façade pair. These "hand in hand" façades are stitched as a near-3D shell of a visual object—the composite entry in the VDic. A composite object can be represented by all the keypoints on the shell and the appearance of the shell. For a given unknown image, we can perform partial matching with the shell to identify the corresponding object.

Both isolated and composite objects are the basic objects in the dictionary and then are assigned a unique identifier. Based on these basic objects, more kinds of visual entries could be generated. The first one is the component object which is a common sub-region shared by several basic objects. Once the component objects are filtered out, they can be used to represent the basic object more compactly. Another kind of visual objects is the abstract entry. In fact, any visual attribute—color, texture, shape or any specific pattern—as long as it presents in object entries, can be defined as an abstract object. In some sense, an abstract object corresponds to a visual concept. It also should be noted that the above three kinds of objects and the abstract object can interpret mutually.

6 Conclusions

General object recognition and image understanding is one of the hardest problems for computer vision and multimedia retrieval. By taking the discovery of general objects in large image set as the pivot, this paper proposes GORIUM to solve the general object recognition and image understanding problem in an unified framework. GORUIM is a four-layer bottom-up model. For the lower two layers, we have proposed and implemented several salient region detection and segmentation algorithms which can precisely extract visual objects from any image. For the third layer, we have also proposed an unsupervised approach to automatically discover general objects from large image set by pairwise matching with local invariant features. On these bases, visual dictionary construction can be easily implemented by using existing matured algorithms. As a consequence, a general object and image understanding machine (short for GORIUMachine) could be turned out in the near future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 90820003, 60973055), and National Basic Research Program of China (Grant No. 2009CB320906). The first author is supported by Program for New Century Excellent Talents in University of Ministry of Education of China. Professor Bernd Girod and his team are appreciated for their constructive comments on GORIUM during the first author's visiting to Stanford University.

References

1 Smeulders A W M, Worring M, Santini S, et al. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell, 2000, 22: 1349–1380
2 Lowe D G. Distinctive image features from scale-invariant keypoints. Int J Comput Vision, 2004, 60: 91–110
3 Bay H, Ess A, Tuytelaars T, et al. SURF: Speeded up robust features. Comput Vis Image Underst, 2008, 110: 346–359
4 Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009. 248–255
5 Biederman I. Recognition-by-components: A theory of human image understanding. Psycho Rev, 1987, 94: 115–147
6 Itti L, Rees G, Tsotsos J. Neurobiology of Attention. San Diego: Elsevier, 2005
7 Li J, Tian Y H, Huang T J, et al. Probabilistic multi-task learning for visual saliency estimation in video. Int J Comput Vision, 2010, 90: 150–165
8 Li J, Tian Y H, Huang T J, et al. Cost-sensitive rank learning from positive and unlabeled data for visual saliency estimation. IEEE Signal Process Lett, 2010, 17: 591–594
9 Li J, Tian Y H, Huang T J, et al. Multi-task rank learning for visual saliency in video. IEEE Trans Circuits Syst Video Technol, 2011, 21: 623–636
10 Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell, 1998, 20: 1254–1259
11 Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009. 1597–1604
12 Hou X, Zhang L. Saliency detection: a spectral residual approach. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, Minnesota, USA, 2007. 1–8
13 Ma Y, Zhang H. Contrast-based image attention analysis by using fuzzy growing. In: Proceedings of the 11th ACM International Conference on Multimedia, Berkeley, CA, USA, 2003. 374–381
14 Yu H N, Li J, Tian Y H, et al. Automatic interesting object extraction from images using complementary saliency maps. In: Proceedings of ACM Multimedia, Firenze, Italy, 2010. 891–894
15 Goferman S, Manor L Z, Tal A. Context-aware saliency detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 2010, 2376–2383
16 Harel J, Koch C, Perona P. Graph-based visual saliency. Adv Neural Inf Process Syst, 2007, 19: 545–552
17 Seo H J, Milanfar P. Static and space-time visual saliency detection by self-resemblance. J Vision, 2009, 9: 1–27
18 Rother C, Kolmogorov V, Blake A. GrabCut-interactive foreground extraction using iterated graph cuts. ACM Trans Graphics, 2004, 23: 309–314
19 Boykov Y, Kolmogorov V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans Pattern Anal Mach Intell, 2004, 23: 1124–1137
20 Movahedi V, Elder J H. Design and perceptual validation of performance measures for salient object segmentation. In: Proceedings of IEEE Workshop on Perceptual Organization in Computer Vision, San Francisco, CA, USA, 2010
21 Chen D, Tsai S, Chandrasekhar V, et al. Inverted index compression for scalable image matching. In: Proceedings of IEEE Data Compression Conference, Snowbird, UT, USA, 2010
22 Chen Z, Duan L Y, Wang C Y, et al. Generating vocabulary for global feature representation towards commerce image retrieval. In: Proceedings of IEEE International Conference Image Processing, Brussels, Belgium, 2011

HUANG TieJun was born in 1970. He received the Ph.D. degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology in 1998. Currently he is a professor at the School of Electrical Engineering and Computer Science of Peking University and the vice director of the National Engineering Laboratory of Video Technology of China. He is supported as New Century Excellent Talents in University by Ministry of Education of China. His research interests include image understanding, video coding, digital libraries and digital rights management. He is a council member of Chinese Institute of Electronics, a senior member of China Computer Federation, a board member of Director of Digital Media Project and an advisory board of IEEE Computing Now.

TIAN YongHong was born in 1975. He received the Ph.D. degree in computer application technology from Institute of Computing Technology, Chinese Academy of Sciences in 2005. Currently he is an associate professor at the National Engineering Laboratory of Video Technology, School of Electrical Engineering and Computer Science of Peking University. His research interests include machine learning and multimedia content analysis, retrieval, and copyright management. He is a senior member of IEEE.

Supporting Information

122011-656-video

The supporting information is available online at info.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.