Published in IET Computer Vision
Received on 2nd June 2013
Revised on 28th October 2013
Accepted on 14th November 2013
doi: 10.1049/iet-cvi.2013.0132

ISSN 1751-9632

Global and local exploitation for saliency using bag-of-words

Zhenzhu Zheng[1], *Yun Zhang*[1], *Luxin Yan*[2]

[1]*High Performance Computing Center, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China*
[2]*Science and Technology on Multi-spectral Information Processing Laboratory, School of Automation, Huazhong University of Science & Technology, Wuhan, People's Republic of China*
E-mail: zhangyun_8851@163.com

Abstract: The guidance of attention helps human vision system to detect objects rapidly. In this study, the authors present a new saliency detection algorithm by using bag-of-words (BOW) representation. The authors regard salient regions as coming from globally rare features and regions locally differ from their surroundings. Our approach consists of three stages: first, calculate global rarity of visual words. A vocabulary, a group of visual words, is generated from the given image and a rarity factor for each visual word is introduced according to its occurrence. Second, calculate local contrast. Representations of local patch are achieved from the histograms of words. Then, local contrast is computed by the difference between the two BOW histograms of a patch and its surroundings. Finally, saliency is measured by the combination of global rarity and local patch contrast. We compare our model with the previous methods on natural images, and experimental results demonstrate good performance of our model and fair consistency with human eye fixations.

1 Introduction

The mechanism of selective visual attention helps us direct our gaze towards an object of interest at the first glance. Hence, saliency models can be a preliminary process for object detection which has been widely studied in the recent years [1–5]. Most traditional methods interpret saliency as 'conspicuous' and they mostly work well in simple scenes with obvious objects (e.g. a red flower among green grasses in Fig. 1*a*). However, in many real-world applications, the situation may not look that easy. Complex scenes usually contain objects which share high similarities with the background, or the background looks cluttered (e.g. a squirrel is on the branches in Fig. 1*b*). Although human can grasp the objects with ease in both simple and complex scenes, many traditional saliency models fail because of certain drawbacks. First, most saliency models are based on 'fixed' contrast detector or feature representation. The most classical model proposed by Itti *et al.* [6] implemented simple red/green and blue/yellow colour contrasts, which works for obvious salient targets (e.g. a red object on a green background). However, human vision system may implement many more subtle colour contrast detectors making it functional even in complex scenes [7]. Still, there are many saliency detection approaches (e.g. [8–10]) adopt sparse coding for image representation which explains what happened in the simple cells of primary visual cortex (*V*1) [11]. The assumption is that an image can be represented in terms of a liner superposition of basis functions. However, these models may result in limited discriminative power for representation. The reason is that large training data are needed in order to gain the basis functions. Morever, the size of representation vectors for the image would be fixed regarding to the number of basis functions. Second, many methods calculated saliency in a single perspective. Some methods [6, 12] calculated the saliency on a local perspective. However, because of the lack of the considerations of global properties of the image, the algorithm would easily miss the salient object or be fooled by local distracters. Meanwhile, some other models [8, 13, 14] calculated the saliency on a global perspective. Their general performances are limited because of the lack of local concerns. However, biological evidence shows that both global and local properties of visual scenes interact in the perception of visual information [15]. This is because *V*1 does not only operate as local spatial function, but also has the capability of global scene organisation [4]. Being different from approaches which regard saliency solely as local contrast, global rarity or local rarity, global and local perspectives were jointly considered in [16, 17] by underlying the idea that salient regions are distinctive with respect to both their local and global surroundings. However, their detection performance still could be improved in terms of local detection and image representation, especially for the complex scene.

In this paper, we propose a saliency model using bag-of-words (BOW) and the saliency is measured in both global and local perspectives. This model uses BOW image representation which may lead to a more flexible and

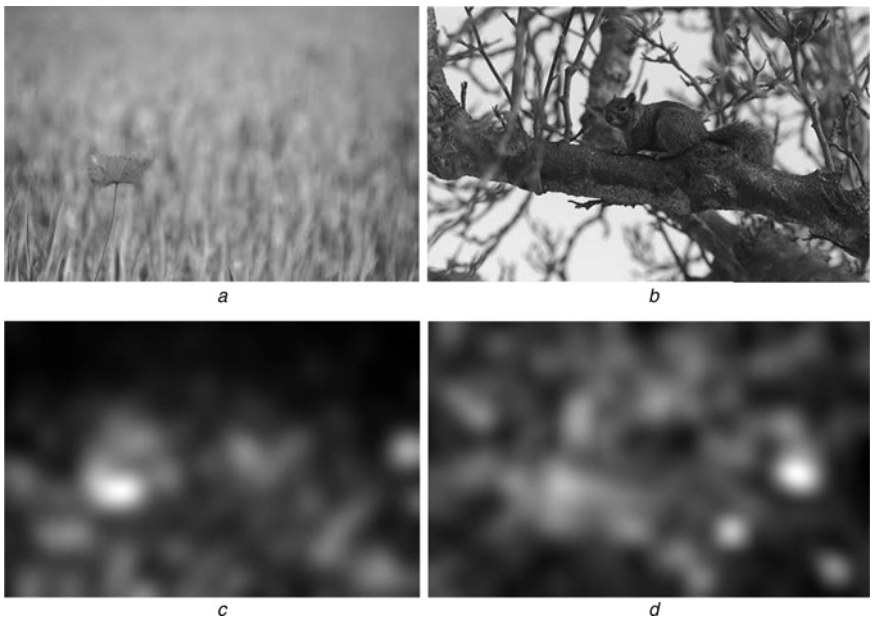

Fig. 1 *Illustration of simple scenes and complex scenes*

a Simple scene with obvious object (the flower);
b Complex scene with non-obvious object (the squirrel) and cluttered background (the branches);
c Itti *et al.*'s [6] saliency detection result of *a* which successfully hit the object (red flower);
d Itti *et al.*'s [6] saliency detection results of *b* which miss the object (the squirrel)

discriminative representation. Besides, saliency is measured in both global and improved local perspectives which make it more consistent with human visual system. Although motivation of both this work and our early one [17] are both built on global and local perspectives, differences lie at the calculations in the local perspective part and representation. In our previous work [17], contrast was measured by the difference between a patch and the entire image. Whereas in this paper we measure contrast by the differences between a patch and its surroundings for better adaptability in complex scenes. In addition, we adopt the BOW representation for the image and its patches for better discriminative power. We regard salient regions as coming from globally rare features and locally contrast regions. We also compare our model with human eye fixation data on viewing natural scenes and apply it in salient object detection in cluttered backgrounds. Experimental results demonstrate good performance of the proposed model.

2 BOW representation

The BOW representation originates from the idea in the text data mining community that an image can be treated as a document and thus 'words' in images can be defined too. In the computer vision community, BOW representation is defined as the histogram representation based on independent features which are known as 'visual words'. Then, the collection of 'visual words' forms a 'vocabulary' [16]. The histogram representation exhibits the proportion of contribution of each visual word in constructing the image. To achieve this, it usually includes following three steps: feature detection, feature representation and vocabulary generation.

For an input image $\boldsymbol{I}$, we generate vocabulary $\boldsymbol{\Omega} = \{W_k\} = \{W_1, \ldots, W_N\}$ through k-mean clustering over the feature sets of image $\boldsymbol{I}$, where $\{W_k\}$ denotes words within vocabulary $\boldsymbol{\Omega}$ and N is the total number of words. Note that this is slightly different from approaches adopting BOW for image classification or object categorisation [18, 19] which usually trains their vocabulary from image datasets. To focus on the given image, we generate one vocabulary for one image towards a more flexible and discriminative representation. For example, Fig. 2 shows two types of images and demonstrates efficiency of BOW representations for an image and its patches. Image A is rich in content which

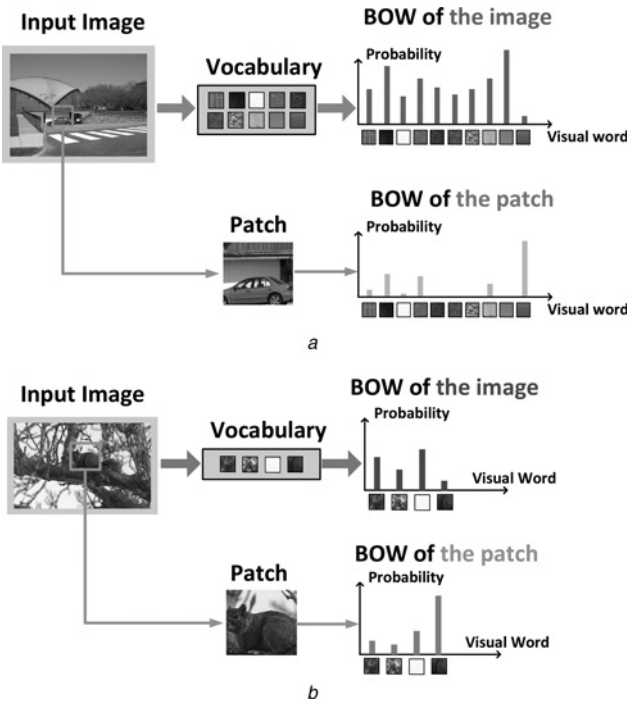

Fig. 2 *Example of BOW representations for an image and its patches. By generating the vocabulary from the given image, the image (patch) representations can be flexible*

a Example image A
b Example image B

doi: 10.1049/iet-cvi.2013.0132

depicts a scene in the park with a gym, a car, sidewalks and trees. While image B conveys less information showing a squirrel on the branch. Using one vocabulary for one image is flexible since visual words that have nothing to do in constructing the scene will be discarded. Moreover, it can be more discriminative. For image B with objects that do not look that conspicuous, with the vocabulary obtained from the given image, we are able to explore even minor feature contrasts.

As vocabulary $\boldsymbol{\Omega}$ is consisted by visual words which together depict the entire image, BOW representation of a local patch $\boldsymbol{I}^m(\boldsymbol{I}^m \in \boldsymbol{I})$ can also be obtained from exactly the same vocabulary $\boldsymbol{\Omega}$. We first detect visual primitives of a patch, and then find out the most appropriate visual word using nearest-neighbour matching. Thus, patch $\boldsymbol{I}^m$ can be denoted as

$$\boldsymbol{I}^m \xrightarrow{\Omega} \boldsymbol{H}^m = \{h_k^m\}_{k=1}^N = \left\{h_1^m, \ ..., \ h_N^m\right\} \tag{1}$$

where notion $\xrightarrow{\Omega}$ indicates the correspondence between a patch and its BOW histogram representation. $\boldsymbol{H}^m$ is a N-dimensional vector representing the histogram of words for patch $\boldsymbol{I}^m$ over vocabulary $\boldsymbol{\Omega}$ with h_i^m indicating the probability of occurrence of word W_i (see Fig. 2).

3 Proposed saliency model

We measure saliency from global and local perspectives and Fig. 3 illustrates the overview of the proposed model. In our approach, saliency is regarded as coming from globally rare features and locally contrasted regions. The proposed model consists of three stages: global perspective, local perspective and saliency estimation.

3.1 Global perspective

The global perspective aims at exploring globally rare features. Inspiration comes from cognitive finding that novel events easily attract human attention [20].

As mentioned in Section 2, the input image $\boldsymbol{I}$ can be denoted by BOW representation $\boldsymbol{H}$ as

$$\boldsymbol{I} \xrightarrow{\Omega} \boldsymbol{H} = \{h_k\}_{k=1}^N = \{h_1, \ \ldots, \ h_N\} \tag{2}$$

where h_k denotes probability of occurrence of visual word W_k. In our implementation, only colour feature is considered. The reason is that human is most sensitive to colour information which has the highest discriminative power among other features such intensity or texture [21]. In our method, RGB colour model is used to represent the input image. This model is an additive colour model where red, green and blue components are added together in various ways to reproduce a broad array of colours.

Here, we introduce a rarity factor r_k associated with each word W_k indicating how anomalous it is in depicting the global image. We define it as

$$r_k = \exp\left(-h_k/\sigma^2\right) \tag{3}$$

The common sense is that low occurrence probability lead to high rarity. Note that we calculated the rarity factor in an exponential function to maintain a balance for rarity factors of all the visual words. Throughout the experiment, sigma for good performance can range from 1.4 to 4 and we set σ to be 3 in this paper.

3.2 Local perspective

The local perspective aims at exploring locally contrasted regions. Following the centre-surround mechanism [20], we calculate the difference between BOW representations of a patch and its surrounding patches.

Let $\boldsymbol{I}^{\mathrm{C}}$ be the centre patch and its surrounding patches together be $\boldsymbol{I}^{\mathrm{S}}$. Similarly with (1), we have

$$\boldsymbol{I}^{\mathrm{C}} \xrightarrow{\Omega} \boldsymbol{H}^{\mathrm{C}} = \{h_k^{\mathrm{C}}\}_{k=1}^N = \left\{h_1^{\mathrm{C}}, \ \ldots, h_N^{\mathrm{C}}\right\} \tag{4}$$

$$\boldsymbol{I}^{\mathrm{S}} \xrightarrow{\Omega} \boldsymbol{H}^{\mathrm{S}} = \{h_k^{\mathrm{S}}\}_{k=1}^N = \left\{h_1^{\mathrm{S}}, \ \ldots, h_N^{\mathrm{S}}\right\} \tag{5}$$

Histogram $\boldsymbol{H}^{\mathrm{C}}$ and $\boldsymbol{H}^{\mathrm{S}}$ denote BOW representations from the centre patch $\boldsymbol{I}^{\mathrm{C}}$ and the surrounding patches $\boldsymbol{I}^{\mathrm{S}}$, respectively. One way to measure difference is to calculate the distance between the two histograms $\boldsymbol{H}^{\mathrm{C}}$ and $\boldsymbol{H}^{\mathrm{S}}$. Here, we use a measurement as χ^2 distance. Thus, local contrast can be

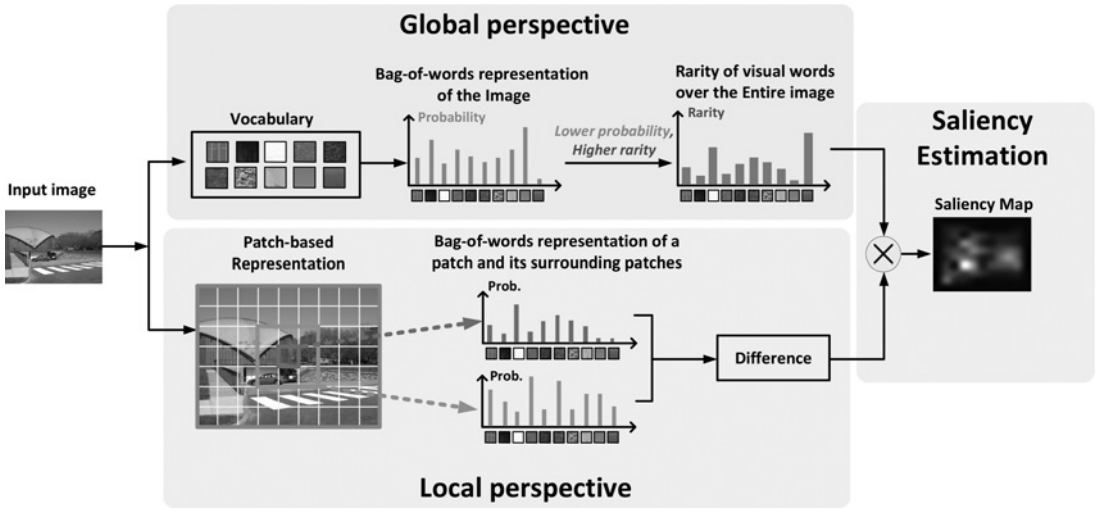

Fig. 3 *Overview of the proposed model*

computed as

$$\text{diff}(\boldsymbol{I}^{\text{C}}, \boldsymbol{I}^{\text{S}}) = \chi^2(\boldsymbol{H}^{\text{C}}, \boldsymbol{H}^{\text{S}}) = \sum_k \frac{\left(h_k^{\text{C}} - h_k^{\text{S}}\right)^2}{(1/2)\left(h_k^{\text{C}} + h_k^{\text{S}}\right)} \quad (6)$$

3.3 Saliency estimation

Finally, saliency is estimated as the combined effect from globally rare features and locally contrasted regions. We measure saliency as the weighted contrast between a patch $\boldsymbol{I}^{\text{C}}$ and its surrounding patches $\boldsymbol{I}^{\text{S}}$, where the weight comes from global features as we intend to highlight influences of rare visual words. Specifically, it is defined as the weighted χ^2 distance between the two BOW histograms $\boldsymbol{H}^{\text{C}}$ and $\boldsymbol{H}^{\text{S}}$

$$\begin{aligned}\text{Saliency}(\boldsymbol{I}^{\text{C}}) &= \text{weighted diff } (\boldsymbol{I}^{\text{C}}, \boldsymbol{I}^{\text{S}}) \\ &= \sum_k \left(r_k \frac{\left(h_k^{\text{C}} - h_k^{\text{S}}\right)^2}{(1/2)\left(h_k^{\text{C}} + h_k^{\text{S}}\right)} \right) \end{aligned} \quad (7)$$

where the first term r_k comes from the global perspective, while the second term $\left(\left(h_k^{\text{C}} - h_k^{\text{S}}\right)^2 / \right)(1/2)\left(h_k^{\text{C}} + h_k^{\text{S}}\right)\Big)$ comes from the local perspective.

4 Experimental results

We evaluate our model by comparison with human eye fixations and application in object detection in cluttered backgrounds.

4.1 Comparison with human eye fixations

We evaluated our model on TORONTO dataset [22] collected by Bruce, which contains eye fixation records from 20 subjects in a viewing task of 120 natural images in size of 681 × 511. We find that the number of visual words may lead to a relatively good performance and this good performance can be maintained with a vocabulary varying from 40 to 100 words. In our experiment, a vocabulary with 60 visual words was generated for each given image. The images were divided into 300 patches equally as much as possible without overlapping. We may slightly change the number of patches for approximation when the image cannot be perfectly equally divided, thus, the patches may have different size. For a single patch, its neighbouring 24 patches will be considered as the surrounding patches. Generally, the number of patches does not depend on the size of an image, but related to the object scale among the image in a certain degree. For example, an image shows a big object, we probably should set a small number of patches. As in this case the object covers major part of an image, only a few patches are necessary to be investigated. On contrary, if we set a large number of patches the size of the patches may become too small and they convey information that is too trivial for depicting an object.

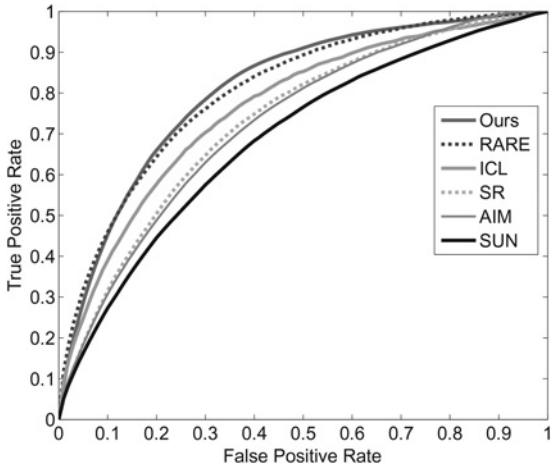

Fig. 4 *ROC curves of our model and other state-of-the-art approaches on the TORONTO dataset*

Table 1 AUCS of different methods

Model	AUC
SUN [12]	0.6884
AIM [8]	0.7257
SR [13]	0.7315
LG [9]	0.7356
ICL [14]	0.7636
RARE [24]	0.8045
ours	0.8104

Similar to the previous works, we adopt AUC (Area Under receiver operating characteristic Curve) as evaluation method with its implementation provided by Judd *et al.* [23] The selection of comparing methods is referred to: citation (Itti *et al.* [6] is widely cited), variety (For computational strategy, Itti *et al.* [6] and SUN [12] are local methods, AIM [8], SR [13] and ICL [14] are global methods. For feature description, Itti *et al.* [6] is the method that fixed contrast features, AIM [8], ICL [14], SUN [12] and LG [9] are methods that use ICA or PCA features), recency (LG [9] and RARE [24] are recent methods). Fig. 4 shows ROC curves of our model and other state-of-the-art approaches. Table 1 lists the AUC of different models and shows that our model achieves the highest performance.

Fig. 5 compares results between the human fixation data and the saliency maps of our model and other state-of-the-art approaches. As shown in the figure , local methods (e.g. Itti and SUN) are easily attracted by the boundary. Global methods (e.g. AIM, SR, ICL) sometimes cause false alarm by regarding non-salient regions as salient ones. LG and RARE are global–local methods. However, LG sometimes miss the salient objects when the input image is complex. Maybe this is because it simply combined local and global saliency. Although our framework is quite similar with that in LG [9], the biggest difference lies at the representation part. The LG adopted a dictionary of 200 basis functions learned from a large repository of natural images, whereas we use a vocabulary (dictionary) learned from the current image. RARE has general good performance with small false alarm being detected. Our model exhibits high consistency with the human visual system because of two reasons. First, we highlight rare features. This is consistent with the basic concept of saliency that novel events are easily attracting human attention [25]. Second, exploiting both global and local properties makes the saliency more similar to human cognitive behaviour [15].

4.2 Salient object detection in cluttered backgrounds

We also tested our model on salient object detection using images collected by Li [26]. The dataset contains 235

doi: 10.1049/iet-cvi.2013.0132

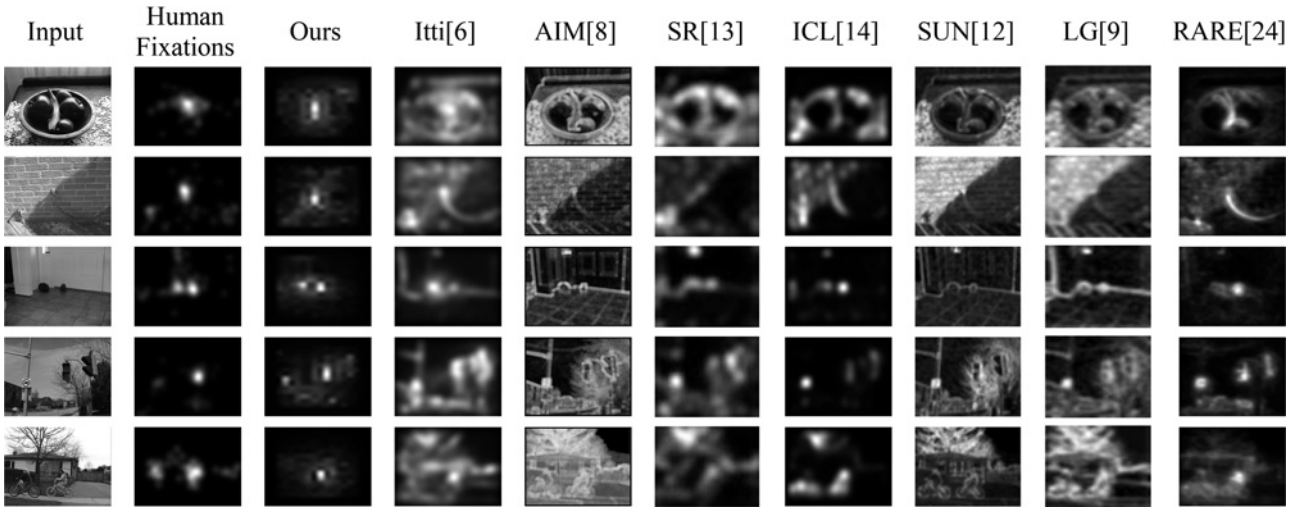

Fig. 5 *Visual comparison of our model and state-of-the-art models over samples from TORONTO dataset*

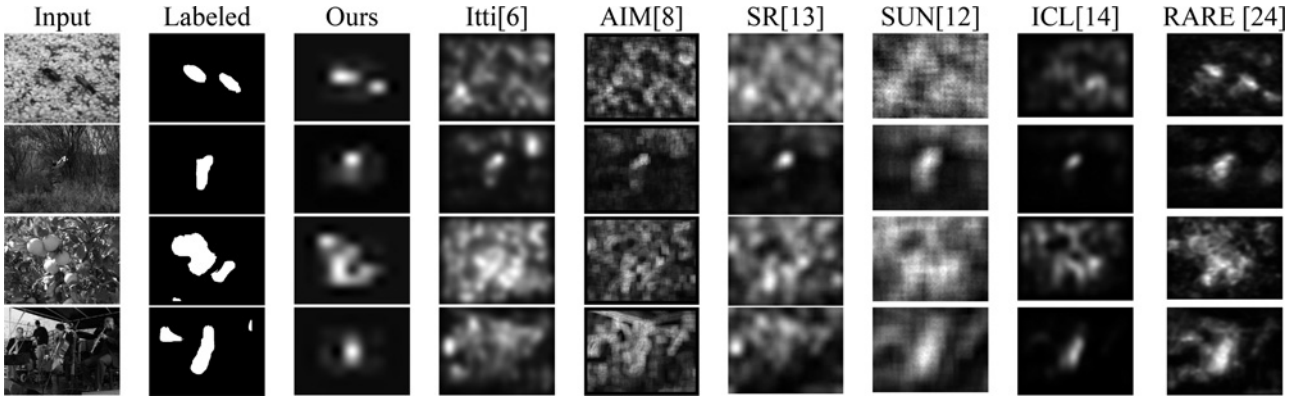

Fig. 6 *Saliency detection results of our model and others detection on images with cluttered background*

colour images in size of 640 × 480 which are divided into six different categories. Both human fixation records (saccades data) and human labelled results are provided in the dataset. In this experiment, we focused on the category (15 images) with cluttered backgrounds to demonstrate the robustness of the proposed model. Ground truth is represented in binary map with one indicating more than half the subjects agreed that the region belonged to a salient object and against for 0. Cluttered background is complex which can be sparse, diverse or with repeated patterns. Saliency detection results are shown in Fig. 6.

Itti *et al.*'s [6] method only implemented simple feature contrasts (e.g. red/green and blue/yellow for colour) which works good for obvious salient targets. However, it may easily fail when objects are not that conspicuous. AIM, SUN and ICL usually fail in cluttered background since they use PCA/ICA features which are trained from a dataset of natural images. The feature representations they use may not remain to be discriminative when the background is cluttered. Our method successfully detects the objects in complex scenes. One reason is we adopt BOW representation from the given image, thus stronger discriminative power can be achieved. The other reason is both improved global and local perspectives have been considered making it more consistent with human perceptive behaviour.

5 Conclusions

In this letter, we present a saliency detection model based on BOW through global and local exploitations. The model effectively conveys the saliency notion of global rare and local contrast. We evaluate the proposed model by the comparison with human eye fixation data and application in object detection task with cluttered background. The experimental results demonstrate good performance of the proposed model. Moving forward, however, our proposed framework can be generalised to handle data such as video and in-depth images.

6 Acknowledgments

This work was supported in part by the the Natural Science Foundation of China under grants 61102088, Shenzhen Emerging Industries of the Strategic Basic Research Project under grant JCYJ20120617151719115 and the Guangdong Nature Science Foundation under grant S2012010008457.

7 References

1 Pal, R., Mukherjee, A., Mitra, P., Mukherjee, J.: 'Modelling visual saliency using degree centrality', *IET Comput. Vis.*, 2010, **4**, pp. 218–229
2 Duncan, K., Sarkar, S.: 'Saliency in images and video: a brief survey', *IET Comput. Vis.*, 2012, **6**, pp. 514–523

3 Liu, Z., Xue, Y., Yan, H., Zhang, Z.: 'Efficient saliency detection based on Gaussian models', *IET Image Process*., 2011, **5**, pp. 122–131
4 Filipe, S., Alexandre, L.A.: 'From the human visual system to the computational models of visual attention: a survey', *Artif.l Intell. Rev.*, 2013, **29**, pp. 1–47
5 Toet, A.: 'Computational versus psychophysical bottom-up image saliency: a comparative evaluation study', *IEEE Trans. Pattern Anal. Mach. Intell*., 2011, **33**, pp. 2131–2146
6 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**, pp. 1254–1259
7 Smirnakis, S.M., Berry, M.J., Warland, D.K., Bialek, W., Meister, M.: 'Adaptation of retinal processing to image contrast and spatial scale', *Nature*, 1997, **386**, pp. 69–73
8 Bruce, N.D.B., Tsotsos, J.K.: 'Saliency based on information maximization', *Adv. Neural Inf. Process. Syst*., 2006, **18**, pp. 155–162
9 Borji, A., Itti, L.: 'Exploiting local and global patch rarities for saliency detection'. IEEE Conf. Comput. Visual Pattern Recognition., 2012, pp. 478–485
10 Wang, W., Wang, Y.Z., Huang, Q.M., Gao, W.: 'Measuring visual saliency by site entropy rate'. IEEE Conf. Computers Visual Pattern Recognition, 2010, pp. 2368–2375
11 Olshausen, B.A., Field, D.J.: 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images', *Nature*, 1996, **381**, pp. 607–609
12 Zhang, L.Y., Tong, M.H., Marks, T.K., Shan, H.H., Cottrell, G.W.: 'SUN: A Bayesian framework for saliency using natural statistics', *J. Vis*., 2008, **8**, pp. 1–20
13 Hou, X.D., Zhang, L.Q.: 'Saliency detection: a spectral residual approach'. IEEE Conf. Computers Visual Pattern Recognition, 2007, pp. 2280–2287
14 Hou, X., Zhang, L.: 'Dynamic visual attention: searching for coding length increments', Proc. 22nd Annual Conf. on Neural Information Processing Systems, 2008, pp. 681–688
15 Forster, J.: 'Local and global cross-modal influences between vision and hearing, tasting, smelling, or touching', *J. Exp. Psychol. Gen*., 2011, **140**, pp. 364–389
16 Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T.: 'A thousand words in a scene', *IEEE Trans. Pattern Anal. Mach. Intell*., 2007, **29**, pp. 1575–89
17 Zheng, Z., Zhang, T., Yan, L.: 'Saliency model for object detection: searching for novel items in the scene', *Opt. Lett.*, 2012, **37**, pp. 1580–1582
18 Wang, G., Zhang, Y., Li, F.: 'Using dependent regions for object categorization in a generative framework'. IEEE Conf. Computers Visual Pattern Recognition., 2006, pp. 1597–1604
19 Li, F., Perona, P.: 'A Bayesian hierarchical model for learning natural scene categories'. IEEE Conf. Computers Visual Pattern Recognition, 2005, pp. 524–531
20 Itti, L., Koch, C.: 'Computational modelling of visual attention', *Nat. Rev. Neurosci*., 2001, **2**, pp. 194–203
21 Nothdurft, H.C.: 'The role of features in preattentive vision: comparison of orientation, motion and color cues', *Vision Res*., 1993, **33**, pp. 1937–1958
22 http://www-sop.inria.fr/members/Neil.Bruce/
23 Judd, T., Ehinger, K., Durand, F., Torralba, A.: 'Learning to predict where humans look'. IEEE Int. Conf. on Computers Vision, 2009, pp. 2106–2113
24 Riche, N., Mancas, M., Gosselin, B., Dutoit, T.: 'Rare: A new bottom-up saliency model'. IEEE Conf. Image Process., 2012, pp. 641–644
25 Ranganath, C., Rainer, G.: 'Neural mechanisms for detecting and remembering novel events', *Nat. Rev. Neurosci*., 2003, **4**, pp. 193–202
26 Li, J., Levine, M.D., An, X., Xu, X., He, H.: 'Visual saliency based on scale-space analysis in the frequency domain', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, pp. 996–1010

doi: 10.1049/iet-cvi.2013.0132