Journal of Systems Engineering and Electronics
Vol. 29, No. 6, December 2018, pp.1300 – 1307



# Real-time object segmentation based on convolutional neural network with saliency optimization for picking

CHEN Jinbo, WANG Zhiheng, and LI Hengyu*

School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072, China

**Abstract:** This paper concerns the problem of object segmentation in real-time for picking system. A region proposal method inspired by human glance based on the convolutional neural network is proposed to select promising regions, allowing more processing is reserved only for these regions. The speed of object segmentation is significantly improved by the region proposal method. By the combination of the region proposal method based on the convolutional neural network and superpixel method, the category and location information can be used to segment objects and image redundancy is significantly reduced. The processing time is reduced considerably by this to achieve the real time. Experiments show that the proposed method can segment the interested target object in real time on an ordinary laptop.

**Keywords:** convolutional neural network, object detection, object segmentation, superpixel, saliency optimization.

**DOI:** 10.21629/JSEE.2018.06.17

## 1. Introduction

The visual categorization and object recognition is an essential task in the computer vision, and it is widely used for content-based image retrieval, car safety, surveillance and the robotics. In works involving physical interactions, object detection and locating are essential components in robotics, which provide challenge problems to the robotics research community that involves object perception, motion planning, grasp planning, and task planning. They are also benchmarks for AI research. The most missing skill within the teams is computer vision by analyzing the Amazon Picking Challenge [1]. Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of the specific class in digital images and videos. For the picking system, it is an essential task to segment target object instance from the environment to provides more accurate localization information of objects.

To recognize the object, many approaches have been proposed such as deformable part detectors [2,3], deep neural networks [4,5] and decision forests [6,7]. In the traditional appearance-based image description approach, objects are often described with visual feature points, by using methods such as scale-invariant feature transform (SIFT) or speeded up robust features (SURF) because of their invariance to the scale, the orientation and almost to the illumination [8]. Recently, convolutional neural networks (CNNs) have been applied to various computer vision tasks such as image classification, semantic segmentation, object detection, and many others. Such great success of CNNs is mostly attributed to their outstanding performance in representing visual data [9]. For the object detection, more recent approaches use region proposal methods to first generate potential bounding boxes in the image and then run the designed classifier on each proposed box. And after classification, post-processing is used to refine the bounding boxes and eliminate the duplicate detections. Though these methods can get the object bounding boxes in the image, the complex pipelines are slow and hard to optimize since each individual component must be trained separately and due to the repeated calculation of the complex classifier, and it does not meet the pixel-level detection demand of the picking system.

To segment object from an image, the traditional template match methods [10,11] match the selected template to image patches for locating and use the contour of the template to segment the object, which subjects to the constraints of the structured environment. Salient object detection methods [12 – 14] exploit the boundary prior, or background information to assist other cues to segment object from image and achieve state-of-the-art results. Due to the absence of high-level knowledge, the category information is absent, and all object is segmented. Recently, convolu-

Manuscript received September 22, 2017.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (61233010; 61305106), the Shanghai Natural Science Foundation (17ZR1409700; 18ZR1415300) and the basic research project of Shanghai Municipal Science and Technology Commission (16JC1400900).

tional networks make progress on the semantic segmentation [15 – 17]. Though the convolutional networks for semantic segmentation achieve the state-of-the-art and are trained end-to-end, they are hard to use because the networks are trained on the pixel-wise label image which is difficult to get. Image segmentation itself is a challenging and unsolved problem, so a hard segmentation not only involves the difficult problem of algorithm selection but also introduces much effort and cost.

Instead of generating potential bounding boxes and then running a classifier on these proposed boxes or semantic segmentation for full image in most recent approaches, perception can be framed as glance and extraction; this is the purpose of our work. Motivated by the fact that humans glance at an image is instantly knowing what objects are and where they are and then focuses on what they are interested in, we propose an architecture consisting of GlanceNet for the glance and saliency optimization for extraction. Redmon et al. [18] provided a solution to frame object bounding box detection as a single regression problem straight from image pixels to bounding box coordinates and class probabilities. Although the detection is completed directly from the full image in one evaluation and is less likely to predict false positives on background, it makes more localization errors. For pixel-wise perception, superpixel algorithms group pixels into perceptually meaningful atomic regions which can be used to replace the rigid structure of the pixel grid. They capture image redundancy, provide a convenient primitive from which to compute image features, and substantially reduce the complexity of subsequent image processing tasks [19].

In this study, we find that the superpixel algorithm can be used for object segmentation but grouping pixels into superpixels for full image is redundant and not necessary; superpixels in promising region are what we actually need, inspired by the human attention. The problem of target object segmentation is transformed to the problem of grouping pixels into superpixels in promising regions and extracting the superpixels that belong to the target; this saves much time in generating superpixels and makes real-time object extraction possible. Finally we use the saliency optimization from robust background detection which achieves object extraction by solving the opposite problem.

## 2. Method

In this section, the pipeline and implementation details of the proposed real-time target object segmentation method are described. The proposed target object segmentation method is outlined in Fig. 1. The method consists of three modules: glance for image understanding, superpixel clustering, and extracting. To find the region of interest quickly and reserve more processing for these promising regions, the main idea of the proposed object segmentation method is to frame image understanding as a single regression problem completed directly from full image in one evaluation firstly. The main purpose of the glance module is to imitate human glance instantly knowing what objects are and where they are so that the redundancy is significantly reduced and more processing can be reserved. The superpixel clustering module groups pixels into perceptually meaningful atomic regions, providing a convenient primitive and reducing the complexity of segmentation. Finally, the superpixels are fed into the extracting module to extract the target object. It is simple to understand and use. Methods with pipeline architecture is easy to generate redundancy, making real-time performance impossible. In this paper, we combine them to reduce redundancy. With the redundancy significantly reduced and flexibility of the proposed architecture, it is possible to be extended to incorporate more complicated method to optimize the performance for specific scenes. The implementation details will be discussed in the following sections.

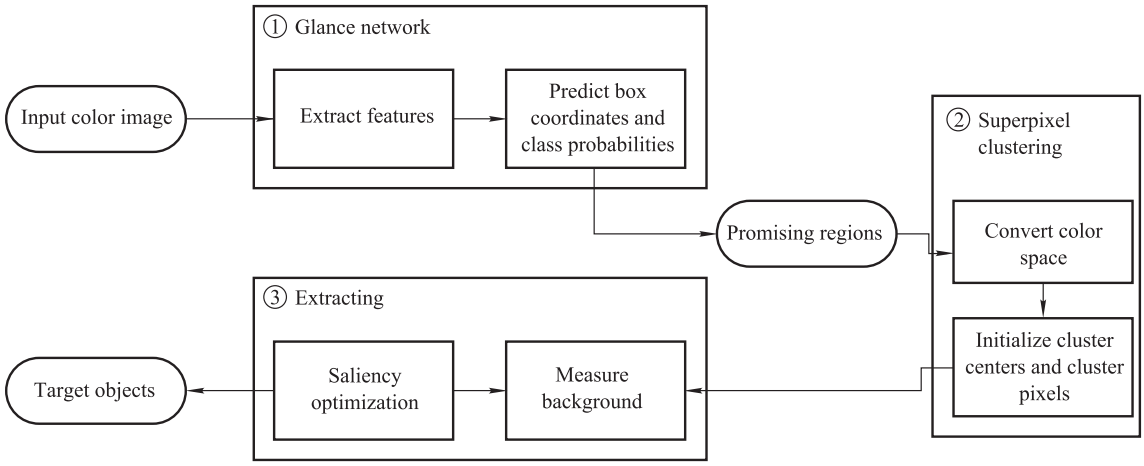



**Fig. 1 Flowchart of the proposed target object segmentation method**

### 2.1 Glance network

To understand the content of image, current detection systems repurpose classifiers to perform detection such as deformable parts models (DPM) using a sliding window approach at even space locations over the entire image, region-based convolutional neural network (R-CNN) running classifier on the boxes generated by the region proposal methods; some systems do semantic segmentation on the image with pixel-wise. Both of them are slow for the complex pipelines and hard to optimize or hard to get training data, so an approach framing object detection as a regression problem is proposed. Redmon et al. [18] provided a model referred to you only look once (YOLO) which can predict bounding boxes and class probabilities directly from a full image in one evaluation. Although it is fast and less likely to predict false positives on background, it makes more localization errors. Inspired by the human's attention mechanism and the idea of YOLO, we frame the target object segmentation as a glance at the image to understand the image for promising regions and then extracting the target object. For the glance module, what we need are understanding of image to know whether the interested target appears and getting the rough location of the target if it appears.

We propose a network referred to GlanceNet named from its original idea illustrated in Fig. 2. It receives a 416×416 color image, and has nine convolutional layers with the first five layers followed by the max pooling layer, and all the convolutional layers without the final layer are combined with the rectified linear units (ReLU) layer for non-linear activation. The first part of network extracts features from the image, while the final convolutional layer with the logistic regression for region prediction and the softmax regression for class prediction. The second part of the network consists of stacked convolution layers with the final $1\times1$ convolution to achieve the detector, which is different from YOLO using stacked full connection layers to achieve the detector. Thus, the improved convolutional layer has the characteristics of local interconnection and shared weights, the ratio of calculations to parameters is higher than that of the full connection layer, so it has high parameter reuse and facilitates parallel acceleration. Otherwise, the proposed network divides the input image into a $13\times13$ grid; each grid cell predicts five bounding boxes and confidence scores for those boxes by the detector glance on the map. As a result of these improvements, GlanceNet has a relatively small decrease in precision while significantly increasing the computational speed, while our GlanceNet uses fewer convolutional layers for feature extraction than YOLO. Convolution is the most common operation in nature, all signal observation, acquisition, transmission; processing can be achieved using the convolution process. In our model, the logistic regression and softmax regression are transformed to logistic and softmax activation on the outputs of the last convolutional layer, for the $1\times1$ convolution can be treated as linear regression. The loss function of this network directly corresponds to the detection performance, and the entire model is trained jointly.

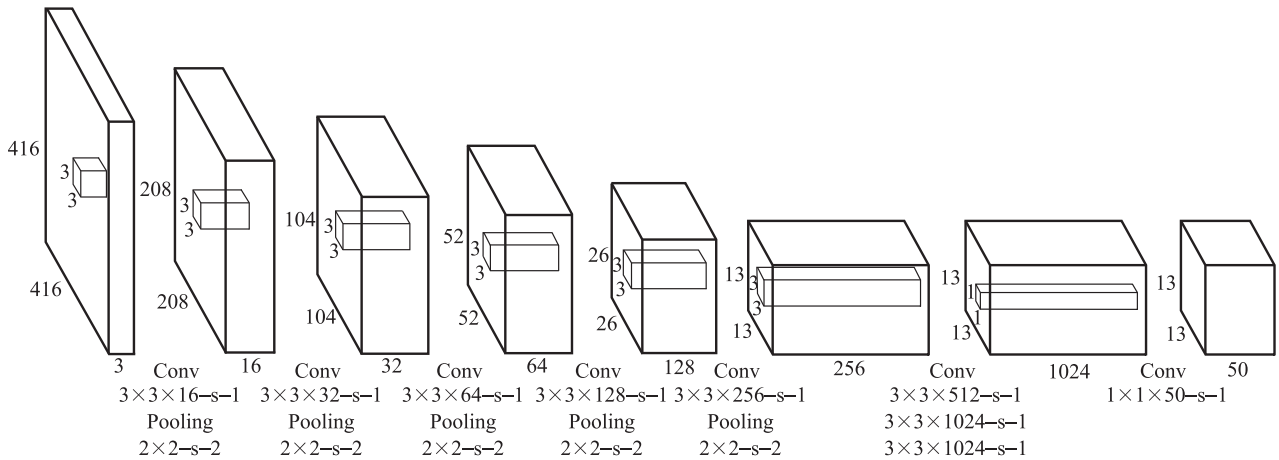



**Fig. 2 Architecture of the GlanceNet**

To train the network, we use the annotation information in PASCAL VOC [20] format and convert the label information to type, center location, width, and height of bounding box before training. Normalization of the information except type is also done to make them fall between 0 and 1. All the annotation information which can be converted to PASCAL VOC format are easy to use when training the model. In the inference process, GlanceNet predicts 845

bounding boxes per image and five class probabilities for each box. In this study, we do not focus on the problem of detection with a large number of categories.

### 2.2 Superpixel segmentation

Superpixels have become an essential tool to the vision community for its distinctive characteristics involve capturing image redundancy, providing a convenient primitive to compute image features, and significantly reducing the complexity of subsequent processing tasks. Achanta et al. [19] presented a superpixel algorithm referred to simple linear iterative clustering (SLIC), which adapts $k$-means clustering to generate superpixels.

SLIC is simple to understand and use which because the only parameter of the algorithm is $K$ by default, the desired number of approximately equally sized superpixels. Firstly, $K$ initial cluster centers $\boldsymbol{Ci} = (l_i, a_i, b_i, x_i, y_i)$ are sampled on a regular gride spaced $S$ pixels apart in CIELAB color space. Secondly, each pixel $i$ is associated with the nearest cluster center, whose search region overlaps location of the pixel, in the assignment step. Once each pixel has been assigned to the nearest cluster center, an update step is done adjusting the cluster centers to be the mean $(l, a, b, x, y)^{\mathrm{T}}$ vector of all the pixels belonging to the cluster. Finally, a post-processing step is done to enforce connectivity by reassigning disjoint pixels to nearby superpixels if it is needed.

In the initialization step, centers are moved to seed locations corresponding to the lowest gradient position in a 3×3 neighborhood to avoid centering a superpixel on edge and to reduce the chance of seeding a superpixel with a noisy pixel. And it must be noticed that simply defining the distance between cluster center and pixel to be the Euclidean distance in $labxy$ space will cause inconsistencies in clustering behavior for different superpixel sizes. Thus it is necessary to normalize color proximity and spatial proximity by their respective maximum distance within a cluster to combine the two distances into a single measure. And the distance measure can be simplified as shown in

$$D = \sqrt{d_c^2 + \left(\frac{d_s}{S}\right)^2 m^2} \tag{1}$$

where $m$ is a constant to normalize color proximity and weigh the relative importance between color similarity and spatial proximity.

### 2.3 Saliency optimization

Salient object detection gets much attention for its importance in applications with low-level cues, especially in the absence of the scene of high-level knowledge. In the applications absence of high-level knowledge, all bottom-up methods rely on assumptions about the properties of objects and backgrounds and exploited the boundary prior and background information to assist other saliency cues such as contrast. It is useful for the object aware image re-targeting, image cropping and object segmentation. Zhu et al. [21] provided a robust background measure method called boundary connectivity which characterizes the spatial layout of image regions concerning image boundaries, and it is much more robust. The basic idea is object and background regions in natural images are quite different in their spatial layout. It assumes that object regions are much less connected to image boundaries than background ones which are consistent with the purpose of our GlanceNet to understand the image and get the rough location of the object. It qualifies how heavily a region $R$ is connected to the image boundaries as shown in

$$BndCon(R) = \frac{|\{p|p \in R, p \in Bnd\}|}{\sqrt{|\{p|p \in R\}|}} \tag{2}$$

where $p$ is an image patch, $R$ is the region to evaluate, and $Bnd$ is the set of image boundary patches.

It is easy to observe that the definition in (2) is intuitive but difficult to compute for the image segmentation itself is a challenging and unsolved problem. To solve the problem, an approach is proposed to group image pixels into superpixels used to replace the rigid structure of the pixel grid, which also captures the image redundancy to reduce the complexity of subsequent processing, and then evaluate the boundary connectivity based on superpixels. By connecting all adjacent superpixels $(p, q)$ and assigning their weight $d_{app}(p, q)$ as the Euclidean distance between their average colors in the CIELAB color space, an undirected weighted graph about superpixels can be constructed. The geodesic distance between any two superpixels $d_{geo}(p, q)$ is defined in

$$\begin{cases} d_{geo}(p,q) = \min\limits_{p_1=p,p_2=,\ldots,p_n=q} \sum\limits_{i=1}^{n-1} d_{app}(p_i, p_{i+1}) \\ d_{geo}(p,p) = 0 \end{cases}. \tag{3}$$

It accumulates the edge weights along their shortest path on the graph. And then the "spanning area" of each superpixel $p$ can be defined in

$$Area(p) = \sum_{i=1}^{N} \exp\left(-\frac{d_{geo}^2(p, p_i)}{2\sigma_{geo}^2}\right) = \sum_{i=1}^{N} S(p, p_i) \tag{4}$$

where $N$ is the number of superpixels, $Area(p)$ computes a soft area of the region that $p$ belongs to and $\sigma_{geo}$ is an adjustment parameter.

Similarly, the length along the boundary is defined in

$$Len_{bnd}(p) = \sum_{i=1}^{N} S(p, p_i) \cdot \delta, \quad p_i \in Bnd \tag{5}$$

where $\delta$ is 1 for superpixels on the image boundary and 0 otherwise.

And we can compute the boundary connectivity using

$$BndCon(p) = \frac{Len_{bnd}(p)}{\sqrt{Area(p)}}. \tag{6}$$

Then the salient object detection problem is transformed to the optimization of the saliency value of each image superpixel. The objective cost function is designed to assign the object region value 1 and the background region value 0. Let the saliency values of $N$ superpixels be $\{s_i\}_{i=1}^{N}$, the cost function can be defined as shown in

$$\text{loss} = \sum_{i=1}^{N} \omega_i^{bg} \cdot s_i^2 + \sum_{i=1}^{N} \omega_i^{fg} \cdot (s_i - 1)^2 + \sum_{i,j} \omega_{ij} (s_i - s_j)^2 \tag{7}$$

where the last smoothness term encourages continuous saliency values. The $\omega_i^{bg}$ and $\omega_i^{fg}$ are related to the background and foreground respectively. The $\omega_{ij}$ is large in flat regions and small at region boundaries. It is defined in

$$\omega_{ij} = \exp\left(-\frac{d_{app}^2(p_i, p_j)}{2\sigma_{clr}^2}\right) + \mu \tag{8}$$

where $\mu$ is a small constant to regularize the optimization in cluttered image regions, and $\sigma_{clr}$ is an adjustment parameter.

## 3. Experimental results

In this section, the proposed real-time target object segmentation method is verified by experiments. The GlanceNet understands the image and then provides the category and location cues to assist other low-level cues to segment object instance from the image.

To verify the GlanceNet, we train and verify the network on the PASCAL VOC dataset using the evaluation metrics widely accepted on the bounding box detection problem. One important evaluation metrics on the detection problem is the intersection of the union (IoU) between ground truth and detection, and the other one is precision-recall curve about the object boxes. In this study speed is also an important evaluation metrics for object extraction is usually time-consuming, so as much as possible processing time should be reserved for the extraction process even if it is not particularly slow. For the intersection of union between ground truth and detection, the bigger it is, the better it is on the detection problem and IoU greater than 0.5 is considered to be correct detection in most benchmarks. We evaluate precision-recall curve on different IoU values because the main purpose of glance is to understand the image, compared with simple detection more localization error can be tolerated.

We train the network with classes chair, bird, bottle, cat, and tv/monitor on 5 540 images and the precision-recall curve is evaluated on 1 484 images. These objects are common in life and difficult to recognize. The intersection of union greater than 0.4 can be considered to be the correct detection; it does not affect the final performance compared to 0.5 for we double the frame before segmentation. The precision-recall curve of our GlanceNet is shown in Fig. 3. And Table 1 shows average precision (AP) of our GlanceNet and other networks. Both the R-CNN [4] and the YOLO use more convolutional layers than our GlanceNet. The DPM [2] and the R-CNN are slow and hard to optimize for the DPM uses a sliding window approach to detect the object, and R-CNN uses selective search to generate potential bounding boxes with convolutional network extracting features on each box.

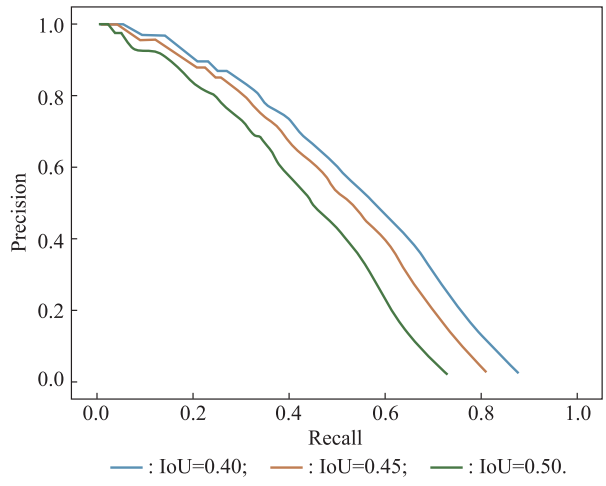



**Fig. 3 Precision-recall curve of GlanceNet on VOC 2007**

**Table 1 Quantitative results on the VOC 2007 dataset**

| Algorithm | AP |
|---|---|
| YOLO | 59.2 |
| R-CNN | 54.2 |
| GlanceNet | 53.3 |
| DPM | 43.2 |

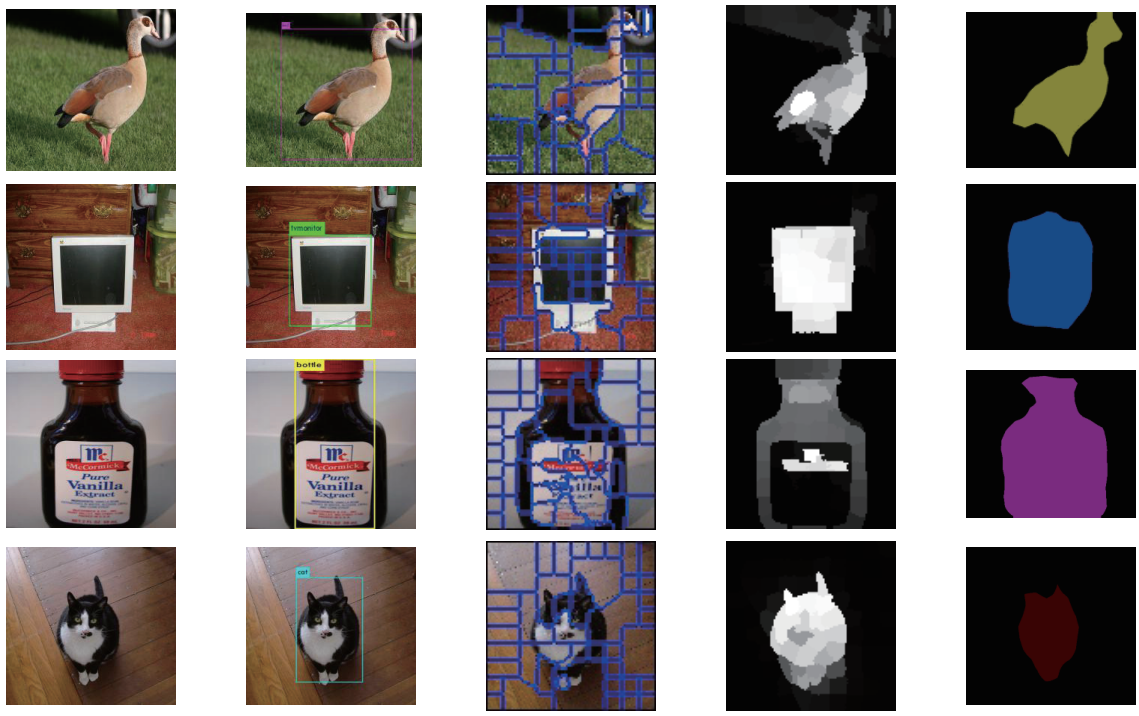
The time spent on each module used in our method and the real-time detector YOLO is shown in Table 2. Our object segmentation method is implemented through C++ programming based on open source computer vision (OpenCV) library, SLIC library, and Caffe [22] framework, and the method is tested on a laptop with a 2.50 GHz central processing unit, 8GB RAM, and Nvidia GeForce GTX 850M. The superpixel algorithm group pixels into perceptually meaningful atomic regions which can be used to replace the rigid structure of the pixel grid after the glance processing. For each region we initiate 100 cluster centers for superpixel cluster in CIELAB color space and the actual super pixels number may be less than 100 because of the forced connectivity step. The superpixel segmentation and saliency optimization are processed in parallel for each region after glance.

**Table 2 Time spent on each module used in our proposed method and YOLO**

| Algorithm | | Time/ms | |
|---|---|---|---|
| YOLO | | 183.96 | |
| Our | GlanceNet | 36.36 | 54.27 |
| | Superpixel | 0.78 | |
| | Saliency optimization | 17.13 | |

For the general 400×300 pixels image, the superpixel segmentation and saliency optimization cost 722 ms on full image, it is hard to optimize and generates more superpixels because it has none prior information of images. In this experiment, we use 100×100 pixels patch for segmentation and saliency optimization which cost 17.9 ms, with prior information we can limit the size of the image and the size of the superpixel.

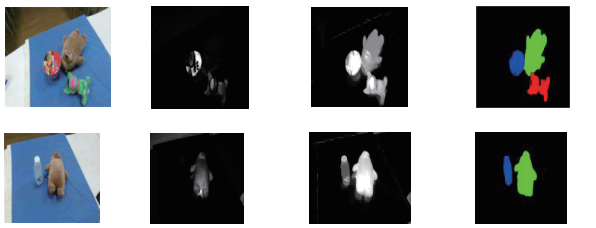
Fig. 4 illustrates the pipeline of our method and the result of FCN-32s [23] for the semantic segmentation. From Fig. 4(b), we can see that the detected frame cannot cover the whole region of the bottle, thus to completely segment the bottle we double the size of the frame, and the whole pixels in the resized frame are used for our saliency optimization based segmentation approach as in Fig. 4(c). The frame part beyond the border of the image is cropped. The resized frame is also beneficial to our saliency optimization using boundary background hypothesis. From the experiment, though the convolutional network for semantic segmentation wants to solve the inherent tension between semantics and location and also reach the state-of-the-art, they are hard to use because of the difficulty to get training data; they are trained end-to-end and pixels-to-pixels. The label images are hard segmentation images which are difficult to get because the image segmentation itself is a challenging and unsolved problem, and the training process for a small amount of training data is a complex tuning and testing process to achieve the desired accuracy.



(a) Input color images (b) Results of GlanceNet (c)Superpixel segmentation of promising regions (d) Optimized saliency maps (e) Semantic segmentation results of FCN-32s

**Fig. 4 Pipeline of our method and result of semantic segmentation method FCN-32s**

Saliency segmentation is a common object segmentation method often used in the traditional object detection which uses the bottom-up approach to detect objects. The results of the state of the art saliency segmentation method saliency filter (SF) [24] and geodesic saliency (GS) [25] are shown in Fig. 5. It can be found that SF and GS segment all the objects together, thus the segmentation method cannot be used for a picking system which requires a segmentation region with a single object; and the image size and superpixel size used for them are difficult to constrain, thus causing high time consuming due to none prior knowledge. Table 3 illustrates the consuming time of the segmentation method SF, GS and the state of the art semantic segmentation method FCN-32s on the general $400\times300$ pixels image. Compared to the saliency segmentation method which segments all the objects together and semantic segmentation which is hard to use because of the pixel-level label image, our method takes the frame detected by the neural network as prior knowledge, the saliency optimization segmentation approach can obtain pixel-level segmentation results for each object accurately and in real time.



(a) Input color images (b) Segmentation results of SF (c) Segmentation results of GS (d) Segmentation results of our method

**Fig. 5 Results of saliency object segmentation**

**Table 3 Consuming time of the saliency object segmentation methods and the FCN-32s method**

| Method | SF | GS | FCN-32s |
|---|---|---|---|
| Time/ms | 336.08 | 271.27 | 56.87 |

## 4. Conclusions

This article has presented a real-time target object segmentation method on RGB color image. The main idea is to frame the perception as the glance for understanding and extraction for object segmentation; global information solves what while local information solves where. This idea has several advantages over previous approaches.

Firstly, compared to the traditional template match methods for object picking which subject to the constraints of structure environment, the proposed method makes progress on general object picking. The GlanceNet understands the image to provide high-level cues to assist salient object detection which has exploited many low-level cues and achieved state-of-the-art results. By the combination of the convolutional network for image understanding and superpixel method, the image redundancy is also significantly reduced. Secondly, more recent approaches use region proposal methods to generate potential bounding boxes in the image and then run classifier on each box for recognition, they are slow and hard to optimize. And in the general object detection based on bounding box, IoU greater than 0.5 is considered to be correct detection which is not enough for picking system. Instead of running classifier on potential bounding boxes which introduces many computing redundancy, we frame the perception as the glance for understanding and extraction for object segmentation which solves the problem of semantics and location for picking system. Thirdly, the training data for the convolutional network is object bounding boxes which are easy to label compared to pixel-wise label image for semantic segmentation.

Kaiming et al. [26] provided a residual learning framework to learn deeper neural networks. The GlanceNet is very simple, without any structure tricks, and it can be further designed with more complexity for the performance more like human glance if using a high-performance computer. And for the superpixel segmentation and saliency optimization, we choose the configures because it is easy to implement. However, it does not perfectly align with the goal of maximizing average precision. The non-fixed number of super pixels can be chosen and a new saliency optimization model can be designed. Our method provides an idea that high-level cues can provide the classes of objects, while low-level cues provide the location of objects, which resolves the inherent tension between the semantics and the location.

## References

[1] CORRELL N, BEKRIS K E, BERENSON D, et al. Analysis and observations from the first amazon picking challenge. IEEE Trans. on Automation Science and Engineering, 2018, 15(1): 172 – 188.

[2] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627 – 1645.

[3] PARK D, RAMANAN D, FOWLKES C. Multi-resolution models for object detection. Proc. of the European Conference on Computer Vision, 2010: 241 – 254.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580 – 587.

[5] OUYANG W, WANG X. Joint deep learning for pedestrian detection. Proc. of the IEEE International Conference on Computer Vision, 2013: 2056 – 2063.

[6] PAISITKRIANGKRAI S, SHEN C, HENGEL A V D. Strengthening the effectiveness of pedestrian detection with spatially pooled features. Proc. of the European Conference on Computer Vision, 2014: 546 – 561.
[7] BENENSON R, MATHIAS M, TUYTELAARS T, et al. Seeking the strongest rigid detector. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3666 – 3673.
[8] HANNAT M, ZRIRA N, RAOUI Y, et al. A fast object recognition and categorization technique for robot grasping using the visual bag of words. Proc. of the International Conference on Multimedia Computing and Systems, 2016: 173 – 178.
[9] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4293 – 4302.
[10] KIRKEGAARD J, MOESLUND T B. Bin-picking based on harmonic shape contexts and graph-based matching. Proc. of the International Conference on Pattern Recognition, 2006: 581 – 584.
[11] FANG J, DENG X, SUN C, et al. A vision based position system for robot picking. Proc. of the International Conference on Electrical and Control Engineering, 2010: 319 – 322.
[12] YANG C, ZHANG L, LU H, et al. Saliency detection via graph-based manifold ranking. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3166 – 3173.
[13] SHEN X, WU Y. A unified approach to salient object detection via low rank matrix recovery. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2012: 853 – 860.
[14] JIANG Z, DAVIS L S. Submodular salient region detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2043 – 2050.
[15] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation. Proc. of the IEEE International Conference on Computer Vision, 2015: 1520 – 1528.
[16] BADRINARAYANAN V, HANDA A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2017, 39(12): 2481 – 2495.
[17] CHEN L C, YANG Y, WANG J, et al. Attention to scale: scale-aware semantic image segmentation. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 3640 – 3649.
[18] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779 – 788.
[19] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274 – 2282.
[20] EVERINGHAM M, GOOL L V, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 2010, 88(2): 303 – 338.
[21] ZHU W, LIANG S, WEI Y, et al. Saliency optimization from robust background detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 2814 – 2821.
[22] JIA Y Q, SHELHAMER E. Caffe: convolutional architecture for fast feature embedding. 2014, arXiv:1408.5093.
[23] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640 – 651.
[24] PERAZZI F, KRÄHENBÜHL P, PRITCH Y, et al. Saliency filters: contrast based filtering for salient region detection. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2012: 733 – 740.
[25] WEI Y, WEN F, ZHU W, et al. Geodesic saliency using background priors. Proc. of the European Conference on Computer Vision, 2017: 29 – 42.
[26] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770 – 778.

## Biographies

**CHEN Jinbo** was born in 1980. He received his M.S. and Ph.D. degrees, both in mechatronic engineering from Shanghai University, Shanghai, China, in 2005 and 2014 respectively. He is currently working as a lecturer in School of Mechatronic Engineering and Automation at Shanghai University. His research interests include computer vision and machine learning.
E-mail: jbchen@shu.edu.cn

**WANG Zhiheng** was born in 1993. In 2015, he enrolled in Shanghai University with a major of mechatronic engineering. He is pursuing his academic master's degree in Shanghai University. He has experienced in some projects about control and vision inspection system. Especially in the machine vision, he has deep research experience. His research interests are in machine vision theory and technology.
E-mail: hengzz@i.shu.edu.cn

**LI Hengyu** was born in 1983. He received his B.S. degree in mechanical engineering and automation from Henan Polytechnic University, China, in 2006, and M.S. and Ph.D. degrees in mechanical and electronic engineering from Shanghai University, China, in 2009 and 2012, respectively. He is currently an associate professor with the School of Mechatronic Engineering and Automation, Shanghai University. His research interests include mechatronics and robot bionic vision system autonomous cooperative control for multiple robots.
E-mail: lihengyu@shu.edu.cn