# QCCE: Quality Constrained Co-saliency Estimation for Common Object Detection

Koteswar Rao Jerripothula [#*§1], Jianfei Cai [*§2], Junsong Yuan [†§3]

[#]*Interdisciplinary Graduate School,* [*]*School of Computer Engineering,* [†]*School of Electrical and Electronic Engineering*
[§]*Nanyang Technological University, Singapore.* email:[1]KOTESWAR001@e.ntu.edu.sg, {[2]ASJFCAI,[3]JSYUAN}@ntu.edu.sg

***Abstract*—Despite recent advances in joint processing of images, sometimes it may not be as effective as single image processing for object discovery problems. In this paper while aiming for common object detection, we attempt to address this problem by proposing a novel QCCE: Quality Constrained Co-saliency Estimation method. The approach here is to iteratively update the saliency maps through co-saliency estimation depending upon quality scores, which indicate the degree of separation of foreground and background likelihoods (the easier the separation, the higher the quality of saliency map). In this way, joint processing is automatically constrained by the quality of saliency maps. Moreover, the proposed method can be applied to both unsupervised and supervised scenarios, unlike other methods which are particularly designed for one scenario only. Experimental results demonstrate superior performance of the proposed method compared to the state-of-the-art methods.**

***Index Terms*—quality, co-saliency, co-localization, bounding-box, propagation, object detection.**

## I. Introduction

Object detection has many applications since it facilitates efficient utilization of computational resources exclusively on the region of interest. Saliency is a common cue used in object detection, but it has only obtained limited success when images have cluttered background. Recent progress in joint processing of images like co-segmentation [1][2][3], co-localization [4], knowledge transfer [5][6][7] has been quite effective in this regard because of the ability to exploit commonness which cannot be done in single image processing.

Despite previous progress, there still exist some major problems for the existing joint processing algorithms. 1) As shown in [1][2], joint processing of images might not perform better than single-image processing for some datasets. This raises up the question: to process jointly or not. 2) Most of the existing high-performance joint processing algorithms are usually complicated due to the way of co-labelling the pixels [2] or co-selection of boxes [4] in a set of images, and also require to tune parameters for effective co-segmentation or co-localization, which becomes much more difficult when dataset becomes increasingly diverse.

There are two types of common object detection: 1) Supervised [6][7], where the task is to populate entire dataset with the help of some available bounding boxes; 2) Unsupervised [4], where the task is to populate entire dataset without any partial labels. In this paper, we handle both types in one framework. Our approach is to iteratively update saliency maps using co-saliency estimation while measuring their quality. For a high-quality saliency map, its foreground and background should be easily separated. Therefore, simple images with a clear background and foreground separation may not need the help from joint processing. For complex images with cluttered backgrounds, by iteratively updating the saliency maps through co-saliency estimation, we are able to gradually improve the saliency maps although they did not have high-quality saliency map to begin with. Images with high-quality saliency maps can play the leading role in the co-saliency estimation of other images. Moreover, some images may already have ground-truth bounding boxes. In such cases, the bounding boxes can replace respective saliency maps as the high-quality saliency maps to help generate better co-saliency maps. Since saliency maps are updated iteratively through co-saliency estimation constrained by their quality scores, we call it **QCCE**: Quality Constrained Co-saliency Estimation. The advantage of such an approach is twofold: (1) It can work effectively for big image dataset and can benefit from high-quality saliency maps; (2) It can automatically choose either the original saliency map or the jointly processed saliency map.

Assuming a Gaussian distribution for foreground and background likelihoods in the saliency map, we make use of the overlap between the two distributions to calculate quality scores. We employ co-saliency estimation [3] along with our quality scores to update the saliency maps. The foreground likelihood of the high-quality saliency map is used as a rough common object segment to define bounding boxes eventually.

Prior to our work, both co-localization [4] problem and bounding box propagation problem [6][7] have been studied on challenging datasets such as ImageNet. While [4] suffers from low accuracy and [6][7] essentially depend upon bounding box availability. In contrast, the proposed method can not only address both problems, but also outperform the existing works.

## II. Proposed Method

Our goal is to define bounding boxes around the common objects in a set of similar images. High-quality saliency maps are obtained while measuring the quality of saliency maps that are iteratively updated via co-saliency estimation. These high-quality saliency maps are then used to eventually define

978-1-4673-7314-2/15/$31.00 ©2015 IEEE 

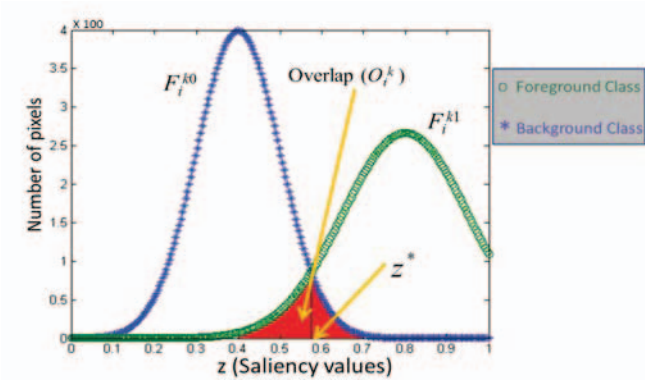


Fig. 1. Quality of saliency map is measured using overlap of estimated distribution of the two classes: Foreground and Background

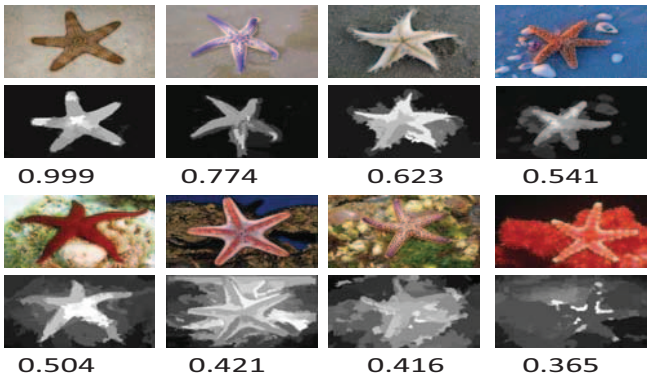


Fig. 2. Sample Images with their saliency maps and quality scores. Saliency maps with low-quality score fail to highlight the starfish.

bounding boxes. In this section, we provide details of quality scores, co-saliency estimation and defining bounding boxes.

### A. *Notation*

Let $\mathbf{I} = \{I_1, I_2, \cdots, I_m\}$ be the image set containing $m$ images and $D_i$ be the pixel domain of $I_i$. Let set of saliency maps be denoted as $\mathbf{S} = \{S_1, S_2, \cdots, S_m\}$ and set of their corresponding quality scores be denoted as $\mathbf{Q} = \{Q_1, Q_2, \cdots, Q_m\}$. For images already having bounding boxes, saliency maps are replaced by respective bounding boxes and their quality scores are set as 1.

### B. *Quality Score*

By quality, we mean how easily two likelihoods (foreground and background) are separable. These likelihoods are formed by thresholding saliency map using the Otsu method. Based on such classification, let $\mu_i^{k1}$, $\mu_i^{k0}$, $\sigma_i^{k1}$ and $\sigma_i^{k0}$ be foreground mean, background mean, foreground standard deviation and background standard deviation for saliency map $S_i$ at the $k^{th}$ iteration (denoted as $S_i^k$), respectively.

Assuming Gaussian distribution for both likelihoods, we denote foreground and background distributions as $F_i^{k1}(z)$ and $F_i^{k0}(z)$, respectively, where $z$ is the saliency value ranging between 0 and 1.

It is clear that the less the two distributions overlap with each other, the better the saliency map is, i.e., the foreground and background are more likely to be separable. In order to calculate the overlap, it is needed to figure out the intersecting point (see Fig. 1). It can be obtained by equating the two functions, i.e. $F_i^{k1}(z) = F_i^{k0}(z)$, which leads to:

$$z^2\left(\frac{1}{(\sigma_i^{k0})^2} - \frac{1}{(\sigma_i^{k1})^2}\right) - 2z\left(\frac{\mu_i^{k0}}{(\sigma_i^{k0})^2} - \frac{\mu_i^{k1}}{(\sigma_i^{k1})^2}\right) + \frac{(\mu_i^{k0})^2}{(\sigma_i^{k0})^2} - \frac{(\mu_i^{k1})^2}{(\sigma_i^{k1})^2} + 2\log(\sigma_i^{k0}) - 2\log(\sigma_i^{k1}) = 0 \quad (1)$$

Let the solution of the above quadratic equation be $z^*$ and the overlap ($O$) can now be computed as

$$O_i^k = \int_{z=0}^{z=z^*} F_i^{k1}(z) + \int_{z=z^*}^{z=1} F_i^{k0}(z) \quad (2)$$

where $O_i^k$ represents the overlap of two classes in $S_i$ at the $k^{th}$ iteration.

Finally, quality score $Q_i$ for $k^{th}$ iteration (denoted as $Q_i^k$) is calculated as,

$$Q_i^k = \frac{1}{1 + \log_{10}(1 + O_i^k)} \quad (3)$$

As we keep updating saliency maps through interaction with other images, we want to choose high-quality saliency maps, i.e. for which maximum quality score is obtained. In Fig. 2, we show a set of images with their saliency maps and quality scores. It can be seen that saliency maps become unfit to highlight the starfish as quality score decreases from top-left to bottom-right.

### C. *Co-saliency Estimation*

The way we update saliency maps after each iteration is through co-saliency estimation which can boost saliency of the common object and suppress background saliency. In order to avoid large variation across images while developing co-saliency maps, in each iteration $k$ we cluster the images into sub-groups by k-means with the weighted GIST feature [2] where saliency maps are used for weights. Let $G_v$ be the set of indexes (i of $I_i$) of images in the $v^{th}$ cluster.

We adopt the idea of co-saliency estimation from [3] where the geometric mean of the saliency map of one image and warped saliency maps of its neighbor images is taken as the co-saliency map. However, we make a slight modification to suit our model, i.e. we use the weighted mean function instead of the geometric mean where weights are our quality scores.

**Saliency Enhancement *via* Warping:** Basically, saliency enhancement takes place at pixel level amongst corresponding pixels. Specifically, following [2], masked Dense SIFT correspondence [8] is used to find corresponding pixels in each image pair. Masks here are the label maps obtained by thresholding the saliency maps. This ensures that pixels having high foreground likelihood play the dominant role in guiding the SIFT flow. The energy function for Dense SIFT flow can

now be represented as

$$E(w_{ij}^k; S_i^k, S_j^k) = \sum_{p \in D_i} \phi\big(S_i^k(p)\big)$$
$$\Big(\phi\big(S_j^k(p + w_{ij}^k(p))\big)||R_i(p) - R_j(p + w_{ij}^k(p))||_1 + \big(1-$$
$$\phi\big(S_j^k(p + w_{ij}^k(p))\big)\big)B_0 + \sum_{q \in N_p^i} \alpha||w_{ij}^k(p) - w_{ij}^k(q)||_2\Big) \quad (4)$$

where $R_i$ is dense SIFT feature descriptor for image $I_i$. The likelihood function $\phi$ for saliency map gives class labels:1 (for foreground likelihood) or 0 (for background likelihood). It can be seen how feature difference is masked by the likelihoods of involved pixels. $B_0$ is a large constant which ensures large cost if the potential corresponding pixel in another image happens to have background likelihood. Weighted by another constant $\alpha$ and likelihood, neighbourhood $N_p^i$ of pixel $p$ is considered for smooth flow field $w_{ij}$ from image $I_i$ to $I_j$.

**Updating Saliency Maps:** Given a pair of images $I_i$ and $I_j$ from a subgroup $G_v$, we form the warped saliency map $U_{ji}$ by $U_{ji}(p) = S_j^k(p')$, where $(p, p')$ is a matched pair in the SIFT flow alignment with relationship $p' = p + w_{ij}(p)$. Since there are quite a few images in subgroup $G_v$, for image $I_i$, we may update its saliency map by computing the weighted mean where weights are respective quality scores, i.e.

$$S_i^{k+1}(p) = \frac{|S_i^k(p)|Q_i^k + \sum_{j \neq i}^{j \in G_v} |U_{ji}^k(p)|Q_j^k}{\sum_{j \in G_v} Q_j^k} \quad (5)$$

This kind of weights ensures that high-quality saliency maps play the leading role in the development of new saliency maps so that new saliency maps evolve towards better ones. Moreover, we also take advantage of prior bounding boxes available which are of high-quality right from the beginning.

### D. *Convergency and Bounding Box Generation*

**Convergency:** If an image reaches its high-quality saliency map, updating of saliency map should stop. Thus, saliency maps of images with ground-truth bounding boxes as high-quality saliency maps get never updated, whereas the saliency maps of other images may get updated depending upon the quality scores. If quality score decreases in next iteration or difference is very small, say 0.005, we will not update the saliency map of the image and proceed for bounding box generation.

**Bounding Box Generation:** Since rough segmentation itself can help developing bounding box, we consider foreground and background likelihoods themselves as common foreground segment and background segment respectively. We get a number of potential sparsely located group of white pixels as objects or connected components using *bwconncomp()* function of MATLAB which we denote as $c$. In order to avoid noisy insignificant objects, we develop an object saliency density metric for each of these objects assuming that real objects would have high foreground saliency density and low background saliency density (here background is rest of the pixel domain). Therefore saliency density metric is defined as:

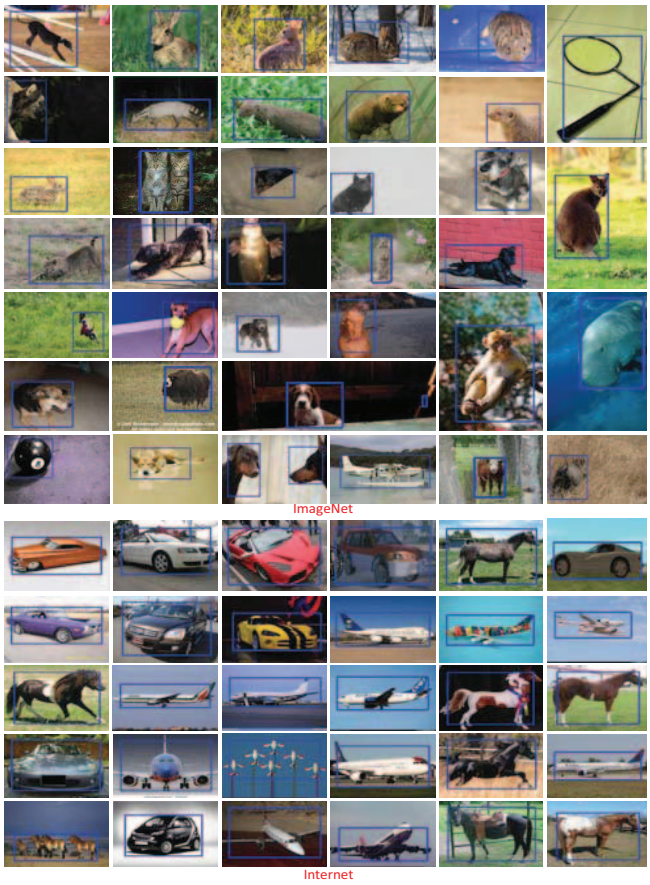


Fig. 3. Sample Results from ImageNet and Internet datasets

$$V(c) = \frac{\sum_{p \in c} S_i(p)}{|c|} - \frac{\sum_{p \in \bar{c}} S_i(p)}{|\bar{c}|} \quad (6)$$

where $\bar{c}$ is the set of the rest of the pixels and $|c|$ is the number of pixels in object $c$. Objects with high saliency density metric are likely to be the real objects. For an image, only those object(s) which are $\geq 50$ percentile according to this object saliency density metric $V$ or is the only object in the image are considered for developing bounding boxes. The bounding box is then drawn using topmost, bottommost, leftmost and rightmost boundary pixels of the qualified objects.

## III. Experimental Results

As per our claim that the proposed method can work better in both unsupervised and supervised scenarios, we use same large scale experimental set up as [4] and [7] for co-localization and bounding box propagation problems, respectively. Following [4], we use CorLoc evaluation metric, i.e., percentage of images that satisfy the condition IOU (intersection over union) defined as $\frac{area(BB_{gt} \cap BB_{pr})}{area((BB_{gt} \cup BB_{pr})} > 0.5$ where $BB_{gt}$ and $BB_{pr}$ are ground-truth and proposed bounding boxes maps, respectively. To distinguish between supervised results and unsupervised results, suffixes (S) and (U) are used, respectively. We use the saliency maps of [9] and [10] in co-localization and bounding box propagation setups,

TABLE I
CORLOC COMPARISON ON IMAGENET AND INTERNET IMAGES DATASET IN CO-LOCALIZATION SETUP

| | ImageNet | Internet |
|---|---|---|
| Baseline using [9](U) | 52.9 | 65.6 |
| [2](U) | - | 75.2 |
| [4](U) | 53.2 | 76.6 |
| Proposed Method(U) | **64.3** | **82.8** |

TABLE II
COMPARISON ON IMAGENET IN BBP SETUP

| | CorLoc |
|---|---|
| Baseline using [10](U) | 64.9 |
| [6](S) | 58.5 |
| [7](S) | 66.5 |
| [7]*(S) | 68.3 |
| Proposed Method(S) | **70.9** |
| Proposed Method(U) | 68.6 |

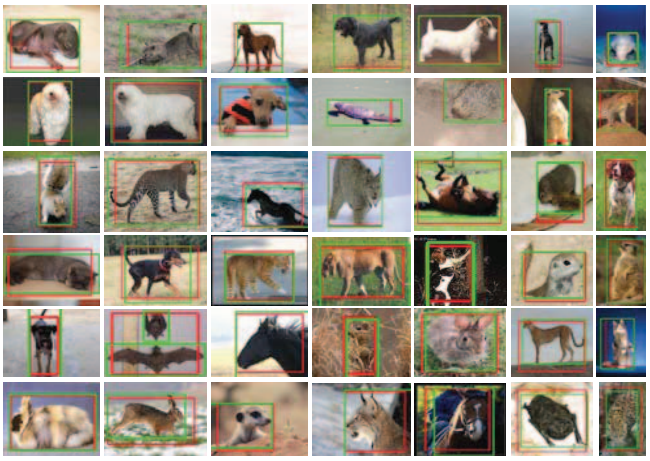

Fig. 4. Sample Visual Comparison between Ground Truth (Green) and Our Results (Red)

respectively. We use bounding boxes generated from these initial saliency maps as baselines.

**Co-localization Setup:** As per the setup in [4], there are 1 million images for which bounding boxes are available on the ImageNet spread over 3627 classes. In addition to ImageNet, for Internet [2] dataset which is actually segmentation dataset, a tight bounding box is developed across each foreground segment and is used as ground truth. We compare our results in Table I for both datasets. It can be seen that we obtain superior results with margin of 11% and 6% improvements on these datasets, respectively. Moreover, our results are obtained without any parameter tuning whereas both [4] and [2] have tuned their parameters on Internet dataset.

**Bounding Box Propagation Setup:** We would like to acknowledge [7] for providing the list of test and training images upon request. In this setup, 32k images are considered for testing purposes, and the saliency maps of the training images of the classes, to which these test images belong, are replaced with bounding boxes. The problem that we are trying to address is similar to "Self" case in [7] where only images within the same class are used for training. In Table II, we compare our results on these 32k test images with two previous attempts in [6][7] to populate ImagNet with bounding boxes in such a supervised manner. [7]* refers to their results using state-of-the-art features and object proposals. Our results are 4% better than state-of-the-art [7]*(S). Considering the proposed method does not essentially need bounding boxes, we report our unsupervised results (Proposed Method(U)) of these 32k test images as well where we do not use any initial bounding boxes and still obtain comparable results to [7]*(S).

Fig. 3 shows sample results obtained on ImageNet and Internet datasets. In addition, we show our results along with ground-truth for visual comparison in Fig. 4. It can be seen that proposed method is able to accurately provide bounding boxes for both simple and complex images because we are able to effectively constrain the joint processing of images.

## IV. CONCLUSION AND FUTURE WORK

We have proposed a QCCE method for common object detection. In the process, we try to address the critical issue of whether to process jointly or not with the help of constraint on the quality of saliency maps. QCCE can act in both supervised and unsupervised ways and obtains superior results in both scenarios. Moreover, it can be well extended to co-segmentation problem. Future work includes incorporating more considerations in the development of quality scores.


## ACKNOWLEDGMENT

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.


## REFERENCES


[1] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 2217–2224.

[2] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 1939–1946.

[3] K. R. Jerripothula, J. Cai, F. Meng, and J. Yuan, "Automatic image Co-Segmentation using geometric mean saliency," in *International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 3282–3286.

[4] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1464–1471.

[5] M. Guillaumin, D. Kttel, and V. Ferrari, "Imagenet auto-annotation with segmentation propagation," *International Journal of Computer Vision*, vol. 110, no. 3, pp. 328–348, 2014.

[6] M. Guillaumin and V. Ferrari, "Large-scale knowledge transfer for object localization in imagenet," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3202–3209.

[7] A. Vezhnevets and V. Ferrari, "Associative embeddings for large-scale knowledge transfer with self-assessment," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1987–1994.

[8] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 5, pp. 978–994, 2011.

[9] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 409–416.

[10] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 2083–2090.