Co-saliency Detection via Base Reconstruction

Xiaochun Cao [1,2], Yupeng Cheng [1,*], Zhiqiang Tao [1], Huazhu Fu [3]

[1] School of Computer Science and Technology, Tianjin University
[2] State Key Laboratory of Information Security, Chinese Academy of Sciences
[3] School of Computer Engineering, Nanyang Technological University
chengyupeng2008@hotmail.com

ABSTRACT

Co-saliency aims at detecting common saliency in a series of images, which is useful for a variety of multimedia applications. In this paper, we address the co-saliency detection to a reconstruction problem: the foreground could be well reconstructed by using the reconstruction bases, which are extracted from each image and have the similar appearances in the feature space. We f rstly obtain a candidate set by measuring the saliency prior of each image. Relevance information among the multiple images is utilized to remove the inaccuracy reconstruction bases. Finally, with the updated reconstruction bases, we rebuild the images and provide the reconstruction error regarded as a negative correlational value in co-saliency measurement. The satisfactory quantitative and qualitative experimental results on two benchmark datasets demonstrate the efficiency and effectiveness of our method.

Categories and Subject Descriptors

I.4.6 [**Image Processing And Computer Vision**]: Segmentation—*Region growing, partitioning*

Keywords

Co-saliency detection; base selection; reconstruction

1. INTRODUCTION

In recent years, increasing researches and technical reports focus on measuring the image saliency, which further proves a signif cant effect of saliency in computer vision [6]. Jacobs *et al.* [7] f rstly propose 'co-saliency' detection searching the unique object in a series of similar images. Extending the concept of saliency to 'multiple' images is a great signif cance, while the requirement of 'common' image narrows the range of its applications. To broaden the usage of co-saliency, later article [5] completes the task in a relaxed def nition, which aims to extracted common salient object from multiple images with no subject to 'similar'. [5] proposes a state-of-the-art cluster-based co-saliency method with three visual attention cues are devised to effectively measure the cluster saliency. Since it learns the global correspondence between a series of images during the clustering process, multiple images co-saliency task can be well completed. However, this method is too sensitive to the clustering result.

Figure 1: Examples of co-saliency detection results. The f rst row is the input images. The second row is saliency map generated by the existing methods [9], and the last row is our co-saliency results.

Recently, a saliency detection method named DSR is provided by [9]. This method makes full use of the visual information of each individual image. It achieves single saliency detection by modelling the simple center bias as a reconstruction problem, where the incompletion of background templates can be well managed. However, DSR [9] is designed for single saliency detection. The center prior is invalid for co-saliency detection from multiple images. And the background template is also invalid in detecting co-saliency from the multiple images which contains the different backgrounds. Moreover, the saliency outliers, which are treated as the single saliency in a few of images, can not be depressed. In this paper, we improve the background templates by using a novel reconstruction bases extracted by the saliency prior and handle the outliers via a selection step.

Fig. 1 shows the comparison between DSR [9] and our method. DSR [9] measures single saliency by background templates obtained from the edge of image. It is weakened in the 'no center prior' cases which frequently exist in co-saliency detection (*e.g.* the person in red at the second image is wrongly depressed). Although the dense reconstruction of DSR [9] can resist this problem by Principal Component Analysis (PCA) to some extend, it also suffers from the situation that large part of foreground exists in the edge of image. As shown in the second row, DSR [9] provides perfect results in the f rst and third images since they are correspond to the center bias. However, when the co-salient foreground appears

*Corresponding author

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2655007 .

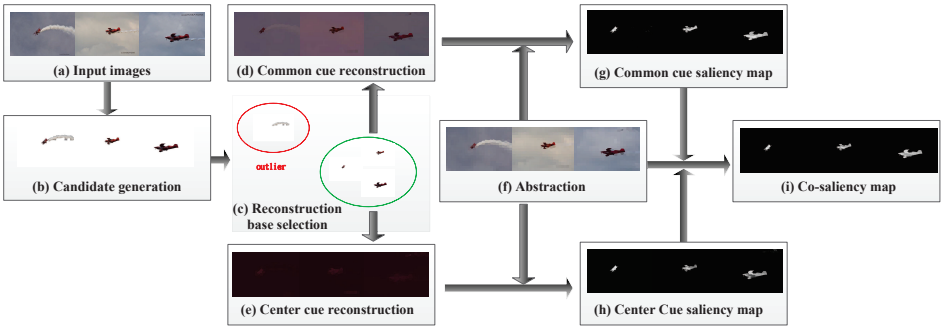


Figure 2: The framework of our reconstruction based co-saliency detection method.

at the edge (the second example) or outlier exists at the center (the last one), the detection result is undesirable as co-saliency. In contrast, our method uses reconstruction base of foreground in place of background templates and utilizes the reconstruction to complement the inaccuracy of single saliency to achieve the co-saliency task. The second and fourth examples illustrate the robustness of our method on complex images.

2. PROPOSED METHOD

Our method is based on the observation that the similarity of co-saliency regions makes them well reconstructed by each other, which is helpful in distinguishing the co-saliency and the background. Based on the def nition of co-saliency, we have the following two roles:

1. The co-saliency is a subset of the single saliency. In other words, the result of single saliency detection method contains the co-saliency region and single saliency appearing in only a part of the image set.

2. Co-salient foregrounds are similar and have close locations in the feature space. This case results in low reconstruction errors with reconstructing the co-salient foregrounds base on each other.

Based on the above roles, we propose a co-saliency detection method via reconstruction error of each image. Fig. 2 shows the framework of our approach. Given a set of images (a), our method generates a candidate set (b), where each image is segmented into a group of superpixels and estimated by various saliency algorithms. We select the reconstruction base (c) by evaluating the occurrence rate on all images. Clusters with low occurrence rate are regarded as outliers and removed from the candidate set (b). Afterwards, we reconstruct each image (d-e) by using common/center cues and compute the reconstruction error comparing with the abstraction of the original image (f), which results in a pair of saliency maps (g-h). Finally, our f nal co-saliency map (i) is generated by gathering the two saliency maps (g-h).

2.1 Candidate generation

We propose the candidate generation step aiming at extracting a rough reconstruction bases via the single saliency detection. Given N input images $\mathcal{I} = \{I^i\}_{i=1}^N$, M saliency detection methods are employed to generate $N \times M$ saliency maps $\{S_j^i\}_{j=1}^M$. To better evaluate the structural information, we f rstly obtain a set of superpixels $\mathcal{X}^i = \{x_p^i\}_{p=1}^{n^i}$ by using method [1] on each image I^i, where n^i denotes the number of superpixels in image I^i. A binary map b_j^i is then computed to represent the voting superpixels by saliency map S_j^i:

$$b_j^i(x_p^i) = \begin{cases} 1, & f_{mean}(S_j^i, x_p^i) \geq \alpha_1 \cdot f_{max}(S_j^i, \mathcal{X}^i) \\ 0, & otherwise \end{cases}, \quad (1)$$

where $f_{mean}(S_j^i, x_p^i)$ computes the mean value of all pixels in superpixel x_p^i by saliency map S_j^i, and $f_{max}(S_j^i, \mathcal{X}^i)$ denotes the max value of all the mean saliency score of $\mathcal{X}^i$. α_1 is a threshold and we set $\alpha_1 = 0.5$ in our experiment. A voting map B^i considering all the saliency map $\{S_j^i\}_{j=1}^M$ is then generated by:

$$B^i = \frac{1}{M} \sum_{j=1}^{M} b_j^i. \quad (2)$$

where B^i has the same size with image I^i and each elements corresponding to a pixel. As all pixels are gathered into a group of superpixels, this map B^i can be explained as how likely a superpixel should be extracted as a candidate. Therefore, a local candidate set f^i of image I^i is def ned as:

$$f^i = d(B^i - \alpha_2 \cdot \mathbf{E}) \cdot I^i, \quad (3)$$

where $\mathbf{E}$ whose all elements are 1 has the same dimension with B^i and α_2 is also a threshold and we set $\alpha_2 = 0.5$ in the experiment. $d(\cdot)$ binaries the input matrix (or value), assign 1 for positive elements, and 0 otherwise. Gathering all local candidate sets of each image, we f nally obtain a candidate set $\mathcal{F} = \{f^i\}_{i=1}^N$ for the input image set $\mathcal{I}$.

2.2 Reconstruction Base selection

Since co-saliency is def ned on the common saliency in all images, the co-saliency bases must be repeated emerge in the candidate set and acquire the advantage of the quantity. Based on this, we then select the reconstruction bases. Given the global reconstruction bases set $\mathcal{F}$, we employ kmeans to cluster all the superpixels into K clusters $C = \{C_k\}_{k=1}^K$. Then a K-bin histogram $\mathbf{z} = \{z_k\}_{k=1}^K$ is created to describe the occurrence rate of all the clusters C in the N input images:

$$z_k = \frac{1}{N} \sum_{i=1}^{N} f(C_k, I^i), \qquad k = 1...K \quad (4)$$

where $f()$ is a binary function, which represents whether the cluster C_k appears in image I^i. If C_k appears in image I^i, $f = 1$, otherwise $f = 0$. At last, the reconstruction base set $\mathcal{G} = \{G_k\}_{k=1}^{n^k}$ is generated by:

$$G_k = \begin{cases} C_k, & z_k \geq \alpha_3 \\ 0, & z_k < \alpha_3 \end{cases}, \qquad k = 1...K \quad (5)$$

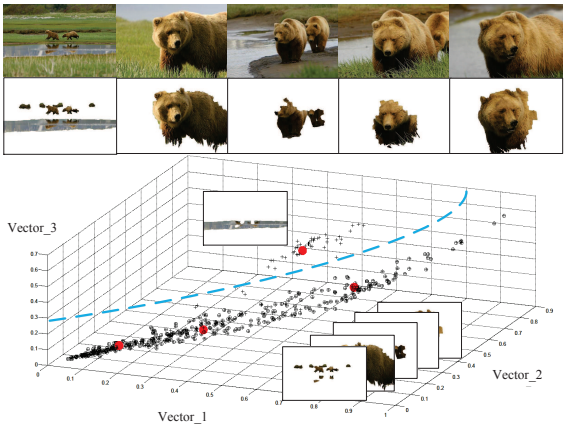


Figure 3: An general view of clustering and f ltrating. The f rst row contains the input images. The second row shows the initial candidate set. The map at bottom is a visual feature space of the f rst three dimensions (*i.e.* RGB color space). Each superpixel in the candidate set is represented by a black cross '+'. Red circles point the clustering center (in this case, clustering number $K = 4$). We mark the selected reconstruction base by $\oplus$. Finally, the selected (up) and unselected (down) superpixels are separated by a blue line.

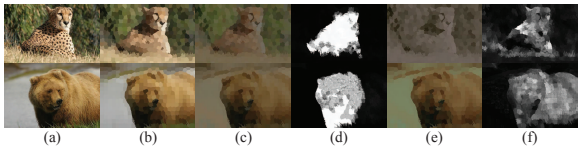


Figure 4: The samples of reconstruction, where the images come from the different co-saliency task. (a) Input images. (b) Abstraction of each image. (c) Common cue reconstruction result. (d) Saliency map based on Common cue. (e) Center cue reconstruction result. (f) Saliency map based on Center cue.

where α_3 is a threshold and we set $\alpha_3 = 0.7$ in our work. Fig. 3 shows an example of base selection in feature space. Depending on the candidate set in second row, the cluster result is shown in a feature space. It is easy to see that the river is wrongly obtained by candidate generation. While, as the river has a different location with the foreground in the feature space, it can be accurately removed from the candidate set.

2.3 Reconstruction error

After obtaining the reconstruction base set $\mathcal{G}$, we map the bases and images $\mathcal{I}$ to an 8-dimension feature space (RGB (3D) and Lab (3D) color, intensity (1D) and the variance of the above 7-dimension in a superpixel) and measure the co-saliency by image construction error following the assumption that there must be a difference between the foreground and background. Each image I^i is then represented as $\mathbf{X}^i = [\mathbf{x}_1^i, \mathbf{x}_2^i, ..., \mathbf{x}_{n^i}^i] \in \mathbb{R}^{8\times N^i}$. We employ two cues to reconstruct the saliency map, common cue and center cue. The outliers is handling by the candidate selection. The missing foreground is covered via center cue. And Common cue is designed for depressing the background.

2.3.1 Common cue reconstruction

Implementing the image reconstruction directly by the reconstruction base set $\mathcal{G}$ is a naive but useful way in measuring the co-saliency. As there are a group of superpixels named reconstruction bases belong to $\mathcal{G}$, we reorder them as $\{\mathbf{l}_h\}_{h=1}^H = \mathcal{G}$ once again for facilitating the description. ($\mathbf{l}_h \in \mathbb{R}^{1\times H}$ and $\mathcal{G} = [\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_H] \in \mathbb{R}^{D\times H}$). To better suppress the background, we employ the reconstruction bases for sparse representation, and encode the image superpixel $\mathbf{x}_p^i$ by:

$$\hat{\beta}_p^i = \underset{\beta_p^i}{argmin} \|\mathbf{x}_p^i - \mathcal{G}\beta_p^i\| + \lambda\|\beta_p^i\|_1, \tag{6}$$

and the reconstruction error of the common cue is:

$$\varepsilon_p^i = \|\mathbf{x}_p^i - \mathcal{G}\hat{\beta}_p^i\|. \tag{7}$$

With the reconstruction error, we could obtain the common cue saliency map (Sec 2.3.3). Fig. 4(c-d) show an example of common cue reconstruction result and saliency map. Since all the reconstruction bases are used in sparse encoding, wood (the f rst row) and river (the second row) in the background are well depressed. However, this approach seriously suffers from the foreground missing of saliency prior. The second row is an obvious example. Body of bear is lost when this part does not exist in reconstruction bases, which results in a high reconstruction error comparing with the head.

2.3.2 Center cue reconstruction

To recover the missing foreground, which is removed by saliency prior, we model the foreground by a group of centers in $\mathcal{G}$. Note that, the centers generated from the f ltrating is an optional choice. However, we can not guarantee that the number of remaining clusters is enough for modelling. Therefore, we extract J center bases $\{\mathbf{U}_j\}_{j=1}^J = \mathcal{U}$ from the reconstruction bases set $\mathcal{G}$ by k-means.

In this way, the foreground is represented by J center base vectors, and each part of foreground is estimated equally no matter whether it constitutes the reconstruction bases $\mathcal{G}$ or not. With the center bases $\mathcal{U}$, we compute the coefficient of superpixel $\mathbf{x}_p^i$ by:

$$\hat{\gamma}_p^i = \underset{\gamma_p^i}{argmin} \|\mathbf{x}_p^i - \mathcal{U}\gamma_p^i\|, \tag{8}$$

where no regularity term is used, since the number of centers is much less than the original reconstruction bases, which can not apply the sparse representation. And the center reconstruction errors is:

$$\eta_p^i = \|\mathbf{x}_p^i - \mathcal{U}\hat{\gamma}_p^i\|. \tag{9}$$

Similar with the common cue, we could generate the center cue saliency map (Sec 2.3.3) via the reconstruction error. Fig. 4(e-f) is center reconstruction result and saliency map. Although the body part is lost in common cue reconstruction, saliency map based on center cue reconstruction cover the foreground by the similarity. Nevertheless, when the background is complex and resemble the foreground (such as the f rst row in Fig. 4(f)), the performance of center cue may be weakened.

In summary, common reconstruction performs well in depressing the awkward background, but sensitive to the saliency prior. Center reconstruction integrates the incomplete foreground while suffers from the complicated background. Hence, the two kinds of reconstruction is complementary to each other and the fusion can achieve a relatively good performance.

2.3.3 Saliency map

We then follow Lu [9] accomplishing the multi-scale to ref ned the error of common/center cues and compute the error of each

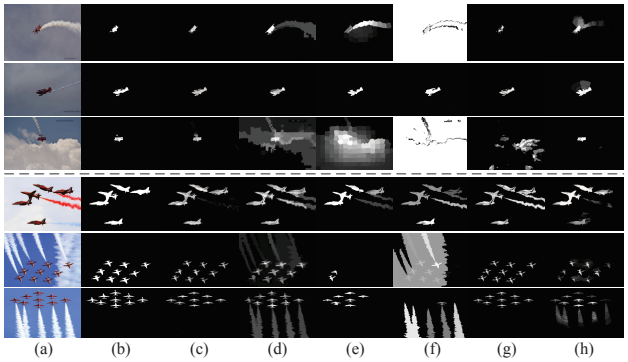


Figure 5: Two examples of saliency detection results. (a) Input images. (b) Ground truth. The example shows a comparison of our algorithm (c) with other saliency methods including (d) RC [4], (e) MR [11], (f) HS [10], (g) FU [5], (h) DSR [9].

pixel z by:

$$E(z) = \frac{\sum_{s=1}^{N_s} \omega_{zn}(s)\nu_n(s)}{\sum_{s=1}^{N_s} \omega_{zn}}, \omega_{zn}(s) = \frac{1}{\|f_z - \mathbf{x}_{n(s)}\|_2}, \quad (10)$$

where N_s is the number of scale, and $\nu_n(s)$ is the error $\hat{\gamma}_p^i$ or $\hat{\beta}_p^i$ in scale n_s. f_z is the feature vector of pixel z, and $n(s)$ denotes the label of superpixel including pixel z in scale s. $\omega_{zn}(s)$ regards the similarity of pixel z with its corresponding superpixel as the weight to average the reconstruction errors in multi-scale. Finally, since the error is negative correlational value in co-saliency measurement the saliency of pixel z is computed by:

$$S(z) = -log(E(z)). \quad (11)$$

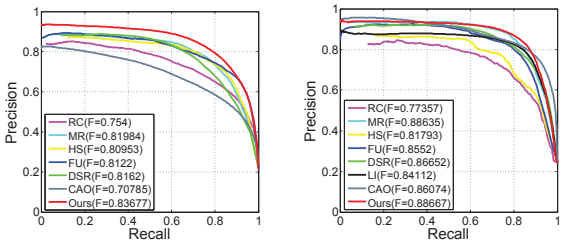
Each reconstruction error (common cue or center cue) produce a saliency result (Fig. 2(g-h)). Gathering both of them, we acquire the final saliency map (Fig. 2(i)).

3. EXPERIMENT

We evaluate our algorithm on two benchmark data sets, and compare our method with seven state-of-the-art saliency detection methods, RC [4], MR [11], HS [10], DSR [9], LI [8] and co-saliency detection methods FU [5] and CAO [3]. (Note that, in this paper, we employ five single saliency methods including MR [11], HS [10], RC [4], DSR [9], and the spatial cue on single saliency in FU [5] to be saliency prior in candidate generation.)

The first experiment tests on the Icoseg dataset [2]. LI [8] is not included since it can only handle a pair of images. Fig. 6(a) shows the Precision-Recall curves of our method and other saliency methods. Our method outperforms all the saliency approaches. The second dataset is Image pair [8]. Fig. 6(b) shows the results and our method overcomes all the other methods on f-measure (**F** beside each method is the f-measure). Nevertheless, because of the simpleness of this data set, each method has a kind of improvement and this situation finally results in a reducing superiority of our approach for the coming of performance ceiling.

Fig. 5 shows some visual saliency detection results. RC [4] fails in low contrast between foreground and outliers. MR [11] and HS [10] achieve high accuracy in some example (second row), but the instability weakens their performance in the overall comparison. DSR [9] still suffers from the center outliers. Although FU [5] well performs in most images, it struggles from the red smog in the forth row. Our method outperforms in the competition and well depress the smog, which is awkward to most algorithms.



(a) co-saliency on Icoseg (b) co-saliency on Image pair

Figure 6: The PR curves of saliency detection on two dataset.

4. CONCLUSIONS

In this paper, we propose a co-saliency detection method based on reconstruction error. The reconstruction base is obtained by the global correspondence selection. With the common and center cues, our method depress the background and handle the missing foreground. The experiments demonstrate our method outperform the state-of-the-art co-saliency methods.

5. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (No. 61332012), National High-tech R&D Program of China (2014BAK11B03), and 100 Talents Programme of The Chinese Academy of Sciences.

6. REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11):2274–2282, 2012.

[2] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. Interactively co-segmentating topically related images with intelligent scribble guidance. *Int. J. Comput. Vision*, 93(3):273–292, 2011.

[3] X. Cao, Z. Tao, B. Zhang, H. Fu, and X. Li. Saliency map fusion based on rank-one constraint. In *International Conference on Multimedia and Expo*, pages 1–6, 2013.

[4] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In *CVPR*, pages 409–416, 2011.

[5] H. Fu, X. Cao, and Z. Tu. Cluster-based co-saliency detection. *TIP*, 22(10):3766–3778, 2013.

[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20(11):1254–1259, 1998.

[7] D. E. Jacobs, D. B. Goldman, and E. Shechtman. Cosaliency: Where people look when comparing images. In *Proc. UIST*, pages 219–228, 2010.

[8] H. Li and K. N. Ngan. A co-saliency model of image pairs. *TIP*, 20(12):3365–3375, 2011.

[9] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, pages 2976–2983, 2013.

[10] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.

[11] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.