SALIENCY MAP FUSION BASED ON RANK-ONE CONSTRAINT

Xiaochun Cao[1,2], *Zhiqiang Tao*[1], *Bao Zhang*[1], *Huazhu Fu*[1], *Xuewei Li*[1]

[1] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
[2] State Key Laboratory Of Information Security, Institute of Information Engineering,
Chinese Academy of Sciences, Beijing 100093, China
caoxiaochun@iie.ac.cn {zqtao, zhangbao, hzfu, lixuewei}@tju.edu.cn

ABSTRACT

Co-saliency is the common saliency existing in multiple images, which keeps consistent in saliency maps. One saliency detection method generates saliency maps for all the input images, so that we have a group of maps. Salient region of each image is extracted by its corresponding saliency map in the group. We use a matrix to combine all the salient regions. Ideally, these co-salient regions are similar and consistent, and therefore the matrix rank appears low. In this paper, we formalize this general consistency criterion as rank-one constraint and propose a consistency energy to measure the approximation degree between matrix rank and one. We combine the single and multiple image saliency maps, and adaptively weight these maps under the rank-one constraint to generate the co-saliency map. Our method is valid for more than two input images and has more robustness than the existing co-saliency methods. Experimental results on benchmark database demonstrate that our method has the satisfactory performance on co-saliency detection.

Index Terms— co-saliency, rank-one constraint, adaptive weight

1. INTRODUCTION

Saliency detection could be considered as a preferential allocation of computational resources [1]. In recent years, many saliency detection methods [2–5] have been designed because of its broad applications, such as object recognition, image segmentation and image database querying. However, most existing saliency models detect the salient object from individual image. We realize that saliency detection in multiple images is different from saliency detection in single image [6–8]. Because the single saliency detection only considers the contrast [2] or uniqueness [3] in one single image. They ignore the relevance information of common saliency in multiple images. Co-saliency is proposed to solve this problem. In this paper, we provide a co-saliency detection method via rank-one constraint.

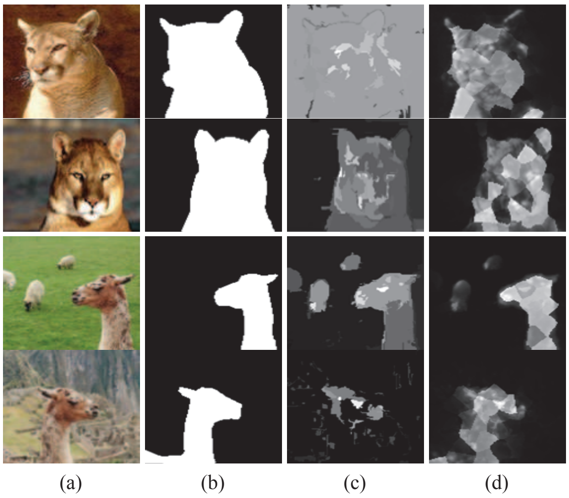


Fig. 1. Visual comparison of single-image saliency map and multiple-single saliency map. (a) Input images. (b) Ground truth. (c) saliency detection [2] in single image. (d) Our co-saliency map.

Figure 1 illustrates the difference between the saliency and co-saliency. The single-image algorithm (e.g. RC [2]) extracts salient object in each image, but it loses the corresponding information of the multiple images. For example, though RC [2] detects the $lion$ in the second row, it fails in the first row because the colors of background and lion are similar. And we also obverse that the $llama$ can not be detected in the last image with the complex background. In contrast, our co-saliency method utilizes the similarity of the objects as additional constraint, and discovers the common object in multiple images.

1.1. Related works

Co-saliency is more useful in various applications, such as co-segmentation [6], common pattern discovery [9], object co-recognition [10]. But few existing methods [6–8] are related to co-saliency detection. Chang *et al.* [6] considers the single-view saliency map and concentrates on those salient parts that frequently repeat in most images. But if the single-view saliency map is invalid, the co-saliency result will degenerate.

Chen *et al*. [7] employs a distribution-based representation to characterize the sparse features in an image. Based on the feature distribution of a pair of images, it presents a progressive algorithm to enhance pre-attentive responses and thus can identify the common salient objects in both images. However, it is hard to generalize this formulation to the case of multiple images.

[8] proposes a method (we call it CSM in this paper) to detect co-saliency of image pairs. The co-saliency detection in [8] is modeled as a linear combination of three SISM maps and two MISM maps. SISM maps are used to describe the region saliency within an image and MISM maps extract the multi-image saliency information from multiple images. In that work, they adopt three types of SISM maps, namely Itti's modle (IT) [1], frequency-tuned saliency (FT) [11], and spectral residual saliency (SR) [3]. And they get two kinds of MISM maps (CC and CP) by respectively matching the color feature and texture feature among the multiple images. If the input images contain a similar region, the region in each image will be given a high saliency value by MISM maps. The weights of these five maps are constant for any pair of images, which MISMs are assigned higher weight values than SISMs. The fixed weights cause a problem that CSM [8] will fail in detecting co-saliency, when MISMs lose effectiveness. Furthermore, CSM [8] only detects the co-saliency for a pair of input images.

In our paper, we propose a general fusion scheme that integrates SISMs and MISMs to co-saliency map. Different from [8], we offer a method to weight these maps self-adaptively, which makes the fusion process more robust. The weight value of each map is decided by the consistence energy with rank-one constraint. We do not only focus on the co-saliency detection, but also care about which maps are more useful to co-saliency detection. Our framework selects the more effective saliency maps in fusion process. Besides, our method is valid for the multiple input images that are more than two.

The rest of our paper is organized as follows. Section 2 gives our method details. Experiments and results are shown in Section 3, followed by conclusion and future work in Section 4.

2. ALGORITHM

As motivated above, our method self-adaptively weights each saliency map participates in fusion process by calculating the rank-one consistency energy. To obtain this energy, we extract the salient regions of the input images by different saliency maps firstly. The regions correspond to the same saliency detection method are placed into one group. We calculate the consistency energy of each group by rank-one constraint, and the energy decides the weights of the saliency maps.

2.1. Saliency cut

In this paper, we choose M saliency detection methods to generate saliency maps for each input image I^i respectively, so each image has M saliency maps S_j^i, $1 \leq i \leq N$, $1 \leq j \leq M$, where N is the number of the input images. Co-salient objects (or regions) in multiple images should be similar. Hence, we judge if one method is contributive to the co-saliency detection by measuring the consistency of all foregrounds (salient regions) that are obtained by it. We divide I^i into two parts:

$$I^i = F_j^i + B_j^i, \tag{1}$$

where F_j^i is the salient region detected by the saliency map S_j^i of image I^i, and B_j^i is regarded as the background of image I^i. Saliency cut method uses saliency map to extract the salient object (or region) from original image. In our method, we are not constrained to specific choice of the segment methods and we obtain F_j^i by identifying salient region via the adaptive threshold [15] as:

$$F_j^i = S_j^i > 2 \times mean(S_j^i). \tag{2}$$

We do saliency cut for all the saliency maps of each image I^i. Each kind of saliency map group $S_j = \{S_j^1, S_j^2, \ldots, S_j^N\}$ is corresponding to a group of salient regions, which is represented by $F_j = \{F_j^1, F_j^2, \ldots, F_j^N\}$. The maps in same group have the same weight, because they are computed by the same method. The weight of S_j is decided by the consistency of the salient region group F_j. To measure it, the region features of all the salient regions in each group must be extracted first.

For each salient region F_j^i in every group F_j, we use histogram to describe its RGB color. Each color channel is evenly divided into 10 bins. Therefore, a $K = 10^3$ dimensions vector is used to describe the histogram. Every pixel in salient region F_j^i is placed into one histogram bin. Pixels located in one same bin are regarded as similar. We sum every bin to form the color histogram for F_j^i, which is denoted as H_j^i. Then we stack N K-dimension horizontal vectors to get the matrix $\mathbf{H_j}$ of F_j, where $\mathbf{H_j} = [H_j^1, H_j^2, \ldots, H_j^N]^T$, $\mathbf{H_j} \in \mathbb{R}^{(N \times K)}$. Note that $\mathbf{H_j}$ is used for measuring the consistency of salient region group F_j, $1 \leq j \leq M$.

2.2. Rank-one constraint and consistency energy

If the co-salient regions in F_j are similar, $\mathbf{H_j}$ ought to be linear correlation, which means any row vector H_j^i can be represented in linear expression by any other vectors in matrix $\mathbf{H_j}$. Hence, the rank of $\mathbf{H_j}$ is asked to be one ideally. But in fact, saliency cut is unable to separate the salient region from original image completely and even the color features of the co-salient regions also have some subtle differences. So, $\mathbf{H_j}$ is a low rank matrix, which is close to one. The rank of $\mathbf{H_j}$ approaches to one more closely with F_j having higher consistency. We call this property as rank-one constraint. An important consequence of this property is that it transforms the consistency measure into measuring the distance between the rank of $\mathbf{H_j}$ and one.

Low rank matrix approximation is an application of singular value decomposition(SVD). To obtain it, we decompose the matrix $\mathbf{H_j}$ by SVD method as:

$$\mathbf{H_j} = \begin{bmatrix} u_1 \cdots u_N \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_N \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_K^T \end{bmatrix} = \sum_{i=1}^{N} \sigma_i A_i, \quad (3)$$

where $A_i = u_i v_i^T$and $\| A_i \|_2 = 1$. Here, we obtain a singular value matrix σ and let σ_i be the singular value on the matrix diagonal in descending order, where $1 \leq i \leq N$. $\tilde{\mathbf{H_j}}^{(\mathbf{k})}$ is the rank k matrix approximation of $\mathbf{H_j}$:

$$\tilde{\mathbf{H_j}}^{(\mathbf{k})} = \sum_{i=1}^{k} \sigma_i A_i. \quad (4)$$

To our best knowledge, $\tilde{\mathbf{H_j}}^{(\mathbf{k})}$ is most close to $\mathbf{H_j}$ in all the rank k matrixes measured by Frobenius norm.

Inspired by the approximation in equation (4), we propose the consistency energy with the rank-one constraint, which is defined as:

$$E_j = \sigma_2 / \sigma_1 \ , \quad (5)$$

where σ_1 and σ_2 denote the first two singular values in equation (3). The matrix rank is equal to the number of the nonzero singular values. If the rank of $\mathbf{H_j}$ is one, the matrix σ will have only one nonzero value σ_1, and $E_j = 0$ makes $\tilde{\mathbf{H_j}}^{(\mathbf{1})}$ = $\mathbf{H_j}$. Otherwise, the rank of $\mathbf{H_j}$ is k ($k \geq 2$). Then the number of nonzero singular values is k and $\tilde{\mathbf{H_j}}^{(\mathbf{k})}$ is close to $\mathbf{H_j}$. Based on rank-one constraint, $\mathbf{H_j}$ is similar to $\tilde{\mathbf{H_j}}^{(\mathbf{1})}$, so that $\tilde{\mathbf{H_j}}^{(\mathbf{k})}$ approximates to $\tilde{\mathbf{H_j}}^{(\mathbf{1})}$. If the σ_2 is small enough to make the other singular values elided, then $\tilde{\mathbf{H_j}}^{(\mathbf{k})}$ will be mainly decided by σ_1. On the other hand, $\tilde{\mathbf{H_j}}^{(\mathbf{k})}$ is not similar to $\tilde{\mathbf{H_j}}^{(\mathbf{1})}$ and F_j should have low consistency. We do not need to know how low it is precisely. We only use the ratio of σ_2 and σ_1 to roughly evaluate the approximation degree between $\tilde{\mathbf{H_j}}^{(\mathbf{k})}$ and $\tilde{\mathbf{H_j}}^{(\mathbf{1})}$, which denotes how close the rank of $\mathbf{H_j}$ is to one. Here, we should notice that our rank-one constraint method is only applicable to the situation of N is small. If N is large, $\mathbf{H_j}$ is more appropriate to be low rank matrix but not one rank matrix.

2.3. Co-saliency assignment

The similarity of the salient region F_j is measured by the consistency energy E_j of the matrix $\mathbf{H_j}$. We make the weight of each saliency map is:

$$w_i = \frac{\exp(-E_i)}{\sum_{j=1}^{M} \exp(-E_j)} \ , \quad (6)$$

where $0 \leq w_i \leq 1$, and $\sum_{i=1}^{M} w_i = 1$. The weight of each saliency map group S_j is determined by the consistency energy of $\mathbf{H_j}$ and it varies with input images. We increase the weight of S_j, which has the low consistency energy. The lower consistency energy implies that the salient regions obtained by the saliency map are similar, and these regions should contribute more in the co-saliency map fusion. Therefore, the consistency energy E_i has high significance and discriminative power. Here, an exponential function is used to emphasize E_i, in order to make the map with lower energy has a higher weight. The co-saliency value of each image in I^i, ($1 \leq i \leq N$) is obtained as:

$$S^i = \sum_{j=1}^{M} w_j \times S_j^i \ . \quad (7)$$

Salient pixels usually gather together in one or some regions in an image. Ideally, the spatial distribution belong to the background will have a high variance, whereas pixels in salient region should be more compact. So we prefer to assign a higher saliency value to the pixels closer to the salient region, and a lower saliency value to the pixels more far away. However, many pixels outside the co-salient regions are salient in one single image. We alleviate the negative effects of these pixels outside the co-saliency regions by taking account of spatial factors in our co-saliency map. Inspired by [4, 12], we reassign the co-saliency map in equation (7) as:

$$S^i(k) = S^i(k) \times \exp(-para \times D(k)) \ . \quad (8)$$

We use the definition of $D(k)$ in [12] and the way to reassignment in [4]. We define the function D(k) as the Euclidean positional distance between pixel k and the closest pixel in co-salient regions F_i, normalized to the range [0,1]. We obtain the F_i by one certain threshold which is set to be 110 as [12]. We use a parameter para to emphasize D(k), and we set para=4 in our experiment according to our experience.

3. RESULT

We choose two MISM maps (CC and CP) which are proposed by [8] and three SISM maps, namely region based contrast saliency (RC) [2], histogram based contrast (HC) [2], and spectral residual saliency (SR) [3]. Total five saliency maps are used to generate the final co-saliency map. In order to compare the different results of different map groups, we also use CC [8], CP [8], SR [3], IT [1] and FT [1] independently to participate in the fusion process. We call the first map group $G1$ and the second $G2$ in our paper.

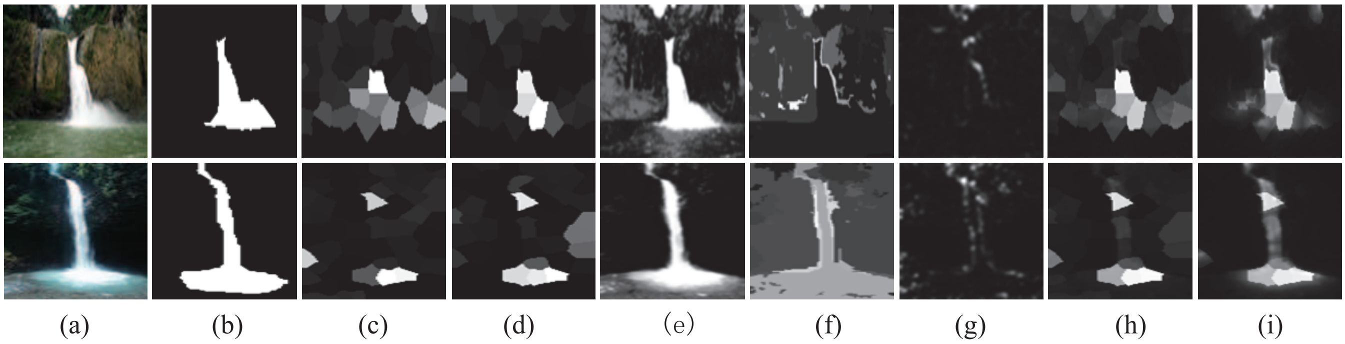


Fig. 2. Visual comparison of saliency maps. (a) Input image pair. (b) Ground truth. (c) CC [8]. (d) CP [8]. (e) HC [2]. (f) RC [2]. (g) SR [3]. (h) CSM [8]. (i) Ours.

Table 1. WEIGHT VALUES ONE IMAGE PAIR

	CC	CP	HC	RC	SR
CSM	0.4	0.4	0.0667	0.0667	0.0667
Ours	0.2181	0.2216	0.2057	0.1695	0.1850

We test our method on a benchmark database of 102 image pairs [8], which ground truth is labeled manually. Each image pair contains one or more similar objects with different backgrounds.

3.1. Qualitative analysis

As we mentioned in Section 1, our method is more robust than [8]. Table 1 shows the weight vales of the maps in Figure 2. We find that the MISM maps mainly decide the final co-saliency map in CSM [8], no matter if they have a good performance on the input image pair. MISM maps emphasize the common regions in multiple images, and they may fail to detect the co-saliency when the input images have the objects with higher intra-class variations, such as the color changes of $waterfall$ in the first column of Figure 2. So the constant weight in CSM [8] may not be the best choice. In Figure 2, HC [2] has detected more consistent salient regions than the other maps. Hence, our method gives a high value to HC [2] in Table 1. Figure 2 shows we have a better visual result than CSM [8] and our result in first row has a more clean background than HC [2].

Our method recalculates weight of every map in $G1$ with the change of the input images, so we call it adaptive weight. We calculate the weight based on the consistency of the salient regions among the input images, which are extracted by the saliency maps S_j. But, when the extracted regions are small part of the correctly co-salient regions, the maps still have a high weight only if the regions are consistent. CC [8] and CP [8] in Figure 2 do not detect all the salient regions and they have a high weight value in Table 1. Although our method has this problem, HC [2] has been given a more higher value by our method than CSM [8]. In practise, we find that our method can not always give the best performance map the highest weight value in our experiment, but we can give it a more higher value than CSM [8]. In other words, we make the weights distribution more suitable.

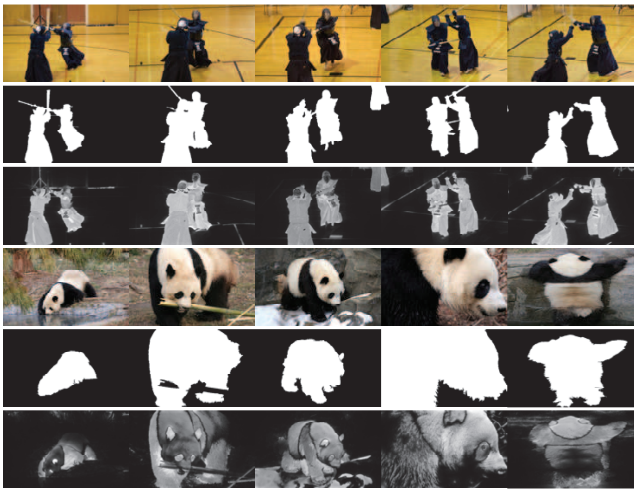
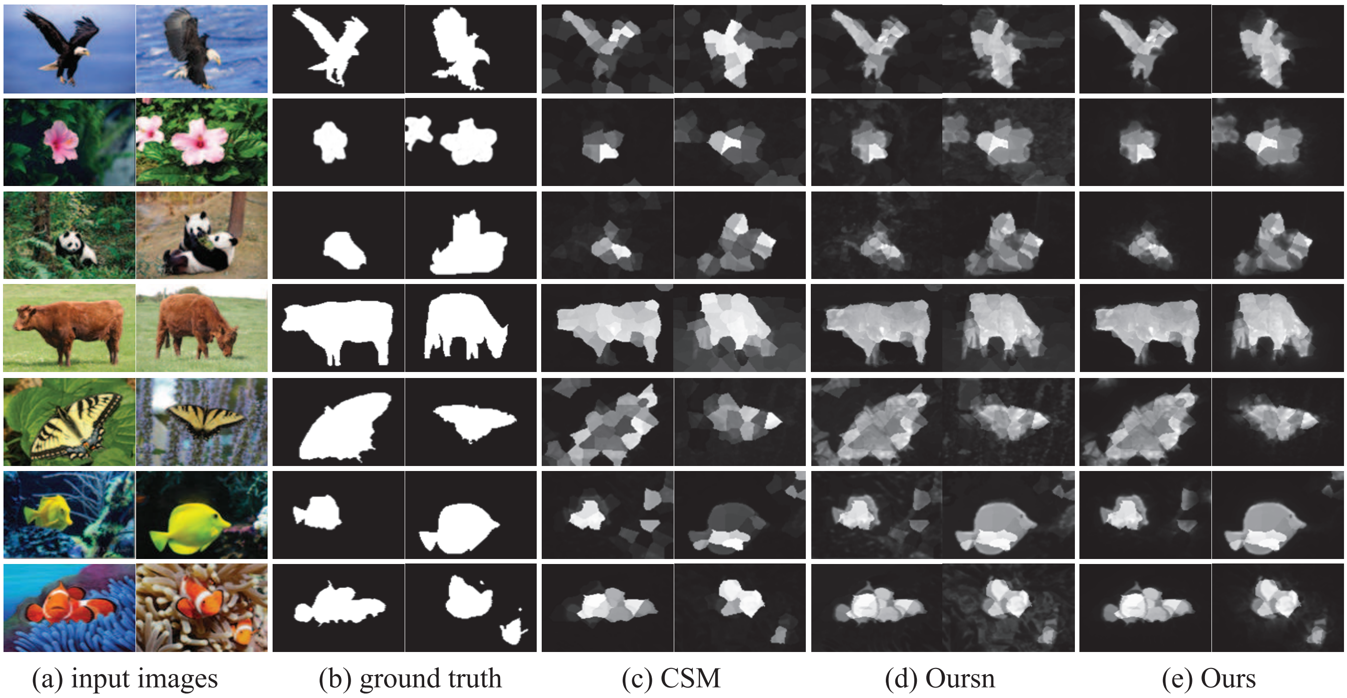


Fig. 3. Visual results of co-saliency detection on benchmark database [13]. The first and fourth rows are the original images. The second and fifth rows are the ground truth and the third and the last rows are our results.

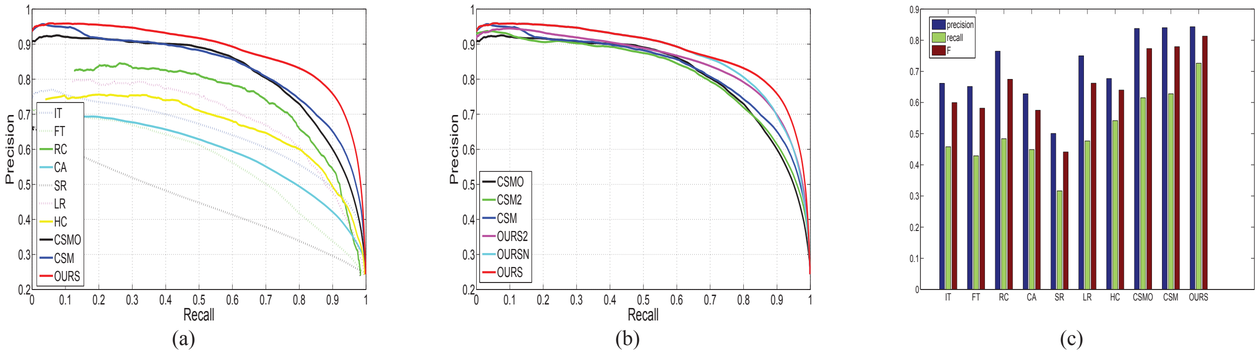
In Figure 4, we show more visual comparisons between our result and [8]. We find that our co-saliency maps in Figure 4 (d), which are not reassigned by equation (8), detect more complete salient regions than CSM [8]. For example, the left $eagle$ in first row and the $flower$ in second row. In addition to this, we separate the salient objects from background more clearly, such as cow in forth row and $butterfly$ in fifth row. Our final maps in Figure 4 (e) can give more details of salient objects and more clean backgrounds in each image pair.

CSM [8] proposes a method to detect the co-saliency for only a pair of images, because CC [8] and CP [8] have the lim-



(a) input images (b) ground truth (c) CSM (d) Oursn (e) Ours

Fig. 4. Visual results of co-saliency detection on benchmark database [8]. (a) Input image pair. (b) Ground truth. (c) Saliency map by CSM [8]. (d) Our saliency map without reassignment by equation (8). (e) Our final saliency map.



(a) (b) (c)

Fig. 5. Precision-recall curves and Precision-recall bars on database [8]. (a) Compared with IT [1], FT [11], RC [2], CA [6], SR [3], LR [5], HC [2], CSMO [8] and CSM [8]. CSMO is the result set which is downloaded from the project page of [8] and CSM is the result with the same method in [8] and the map group G_1. (b) CSM2 and CSM are obtained by the method in [8] with G_2 and G_1 respectively. OURS2 and OURS are with G_2 and G_1 respectively. OURSN is the OURS without reassignment by equation (8). (c) Precision-recall bars for adaptive thresholds.

itation of the input image number. In our framework, the common saliency is highlighted by given a high weight. Hence, we do not have the restriction of the number of the input images. Figure 3 shows our method extracts the co-saliency among the five input images precisely. The co-saliency map is combined with five single-saliency maps, which are RC [2], HC [2], SR [3], FT [11] and LC [14]. The input images come from the database which is offered by [13].

3.2. Quantitative analysis

In order to compare our method to other methods objectively, we follow the two evaluation strategies [11]. In our first experiment, $Precision$ and $Recall$ are calculated by a fixed threshold. We vary the threshold from 0 to 255 to obtain 256 different precision-recall pairs. We compare our method to some state-of-the-art approaches, including contrast-based (IT [1], RC [2], HC [2]), uniqueness-based (SR [3], FT [11]), low-rank (LR [5]), and some other co-saliency approaches (CA [6], CSM [8]). Figure 5(a) gives the ROC curves of all the tested method. Our method achieves the highest precision for most of recall values, which demonstrates we have the best performance.

Figure 5(b) shows we find that the result of our method with $G1$ is better than the result with $G2$. When we compare these two map groups with CSM [8], we have the same result. It seems that RC [2] and HC [2] are more appropriate to co-saliency detection than IT [1] and FT [11] based on this comparison. Therefore, our method provides a kind of evaluation criteria to judge if the map is useful for co-saliency. On the other hand, RC [2] and HC [2] both have higher curves than IT [1] and FT [11] in Figure 5(a), so we may improve co-saliency detection effect by using better saliency map. Figure 5(b) also shows that we can have better result by reassigning the co-saliency map with the equation (8).

In our second experiment, we use the adaptive threshold [11] which is defined as twice the mean saliency of the image to extract the salient regions. $F - measure$ [11] is as:

$$F_r = \frac{(1+\gamma^2)Precision \times Recall}{\gamma^2 \times Precision + Recall}, \tag{9}$$

where $Precision$ and $Recall$ are calculated by the regions whose values are above the adaptive threshold and γ^2 is a positive constant which weights the precision over recall. Here we set $\gamma^2 = 0.3$ as in work [11]. The comparison is shown in Figure 5(c), which indicates that our method has the highest value on $Precision$, $Recall$ and $F - measure$. In fact, we have improved 0.3% , 9.83% and 3.36% compared to CSM [8].

4. CONCLUSION

In this paper, we have presented a robust and effective fusion method to detect the co-saliency among multiple input images. We adaptively weight the saliency maps to generate the co-saliency map of each input image. To obtain the weight, rank-one constraint is proposed to describe the consistency of the salient regions, which are extracted by one map group. We also offer one simple way to compute the consistency energy, which is used to describe the rank-one constraint.

In the future, we plan to use more visual features to calculate the consistency among the input images, and add more saliency maps to the fusion process. Besides, we desire to make CC [8] and CP [8] available for arbitrary input images and extend our algorithm to other visual applications.

5. ACKNOWLEGEMENT

This work was supported by National High-tech R&D Program of China (2012AA011503), National Natural Science Foundation of China (No. 61003200), Tianjin Key Technologies R&D program (No.11ZCKFGX00800), and 100 Talents Programme of The Chinese Academy of Sciences.

6. REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

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

[3] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” *CVPR*, pp. 1–8, 2007.

[4] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” *CVPR*, pp. 733–740, 2012.

[5] Xiaohui Shen and Ying Wu, “A unified approach to salient object detection via low rank matrix recovery,” *CVPR*, pp. 853–860, 2012.

[6] K. Chang, T. Liu, and S. Lai, “From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model,” *CVPR*, pp. 2129–2136, 2011.

[7] H. Chen, “Preattentive co-saliency detection,” *IEEE International Conference on Image Processing (ICIP)*, pp. 1117–1120, 2010.

[8] H. Li and K. Ngan, “A co-saliency model of image pairs,” *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, 2011.

[9] J. Yuan and Y. Wu, “Spatial random partition for common visual pattern discovery,” *ICCV*, pp. 1–8, 2007.

[10] M. Cho, Y. Shin, and K. Lee, “Co-recognition of image pairs by data-driven monte carlo image exploration,” *ECCV*, pp. 144–157, 2008.

[11] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” *CVPR*, pp. 1597–1604, 2009.

[12] S. Goferman, L. Zelnik-Manor, and A.Tal, “Context-aware saliency detection,” *CVPR*, pp. 2376–2383, 2010.

[13] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, “Interactively co-segmentating topically related images with intelligent scribble guidance,” *International Journal of Computer Vision*, pp. 272–292, 2011.

[14] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” *ACM Multimedia*, pp. 815–824, 2006.

[15] D. E. Jacob, D. B Goldman, and E. Shechtman, “Cosaliency: Where people look when comparing images,” *Proc. ACM Symposium on User Interface Software and Technology*, 2010.

[16] Jinqiao Wang, Lei Xu, Hanqing Lu, and Changsheng Xu, “Context saliency based image summarization,” *International Conference on Multimedia and Expo (ICME)*, pp. 270–273, 2009.