Neurocomputing 275 (2018) 1650–1661

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Unsupervised image co-segmentation via guidance of simple images☆

Lina Li[a,b], Zhi Liu[a,c,*], Jian Zhang[b]

[a] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[b] Global Big Data Technologies Centre, School of Electrical & Data Engineering, University of Technology Sydney, Sydney, Broadway NSW 2007, Australia
[c] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China

ARTICLE INFO

Article history:
Received 14 April 2016
Revised 15 April 2017
Accepted 3 October 2017
Available online 6 October 2017

Communicated by Xianbin Cao

Keywords:
Co-segmentation
Image ranking
Image segmentation
Simple image guidance

ABSTRACT

This paper proposes a novel image co-segmentation method, which aims to segment the common objects in a group of images. The proposed method takes advantages of the reliability of simple images and successfully improves the performance. The images are first ranked by the complexities based on their saliency maps. Then, the simple images, in which objects are common and easy to be segmented, are selected and processed to obtain their segmentation results, these segmentation results are taken as the samples of the targeted objects. Finally, the remaining complicated images are segmented with the guidance of the samples. The experiments on the iCoseg dataset demonstrate the outperformance and robustness of the proposed method.

1. Introduction

Image co-segmentation has received a boosting attention by researchers who work on object segmentation. Object segmentation is a fundamental problem in computer vision for many applications such as content-based image retrieval [1,2], object-aware image/video retargeting [3,4], image editing [5,6] and content-aware image/video compression [7,8]. Object segmentation in a single image usually incorporates different measures at the level of pixels [9–11], or at the level of regions [12] into a global objective function; by minimizing the objective function, the objects can be separated from the background. This kind of methods segments the object in each image in isolation, while co-segmentation aims to segment objects with additional information from other images containing the same or similar objects[13]. Thus, image co-segmentation can be derived as a task of jointly segmenting the common objects in a group of images.

Most of the co-segmentation methods are unsupervised and have various problem formulations [14–23]. In [16], the segmentation problem is handled in a discriminative clustering framework, in which objects are separated from the background by merging the image pixels into two clusters that can be maximally distinguished. Disturbance from consistent background is considered in [14], and a framework of saliency detection and discriminative learning is designed to overcome the disturbance. In [15], each image is treated as a player, for which a constrained utility function is designed by using the cooperative game, heat diffusion, and image saliency. Common objects can be segmented according to the labels defined via the constrained utility function. In [17], pixel-level correspondences between images are considered, while in [20,21], it is region-level affinities that are calculated. In [21], the affinities are calculated to score the multiple overlapping segment candidates, then these candidates and their scores are used to generate co-segmentation likelihood maps. In [20], region-level affinities are utilized for each superpixel to find its correspondingly matched superpixels. Some co-segmentation methods utilize correspondences among all the images [19] while others compute correspondences among specific selected images [17]. In [18], a feature pool is built based on overlapped rectangle regions and grows with the number of processed images, and the image to be processed obtains its prior information via the feature pool. In [22], a feature adaptive co-segmentation method is proposed to replace the fixed region similarity measurement for the better region matching. This method first evaluates the image complexities, then selects simple images and obtains the initial segments. This method uses a linear combination of the common features, and the linear combination parameters are learnt from these segments. The learnt adap-

☆ This work was supported by National Natural Science Foundation of China under Grant No. 61471230, and by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

* Corresponding author.
E-mail addresses: Lina.Li-1@student.uts.edu.au (L. Li), liuzhisjtu@163.com (Z. Liu), Jian.Zhang@uts.edu.au (J. Zhang).

https://doi.org/10.1016/j.neucom.2017.10.002
0925-2312/

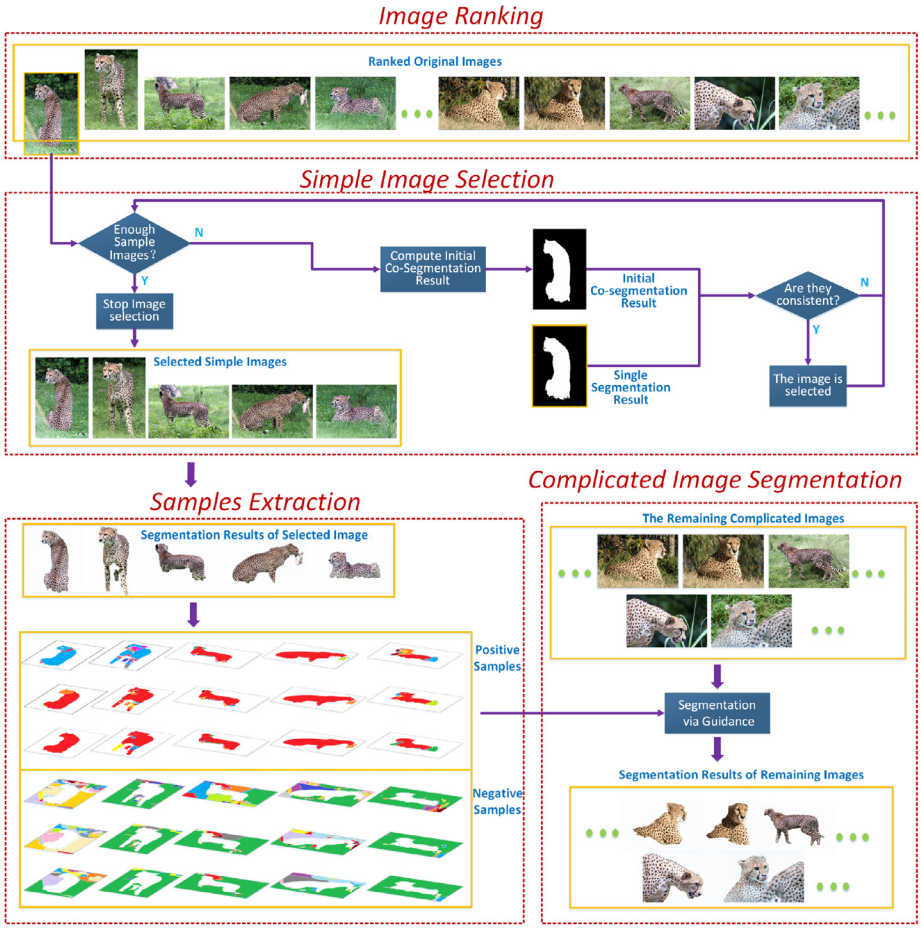


Fig. 1. The flowchart. There are four steps in the proposed method: image ranking, simple image selection, samples extraction and complicated image segmentation.

tive feature is used to segment all the images in an image group. In [23], a saliency boosting model is proposed to optimize the joint segmentation results and saliency maps mutually.

Another kind of co-segmentation methods is interactive co-segmentation that allows a user to decide what the foreground is [24]. This kind of methods may have simpler and reasonable energy functions by keeping a user in the loop, however, keeping a user involved is not always available for some applications.

Image co-segmentation has been extended from two-class segmentation to multi-class segmentation recently [25–28]. In [25], an energy based formulation of multi-class image co-segmentation is proposed which admits a probabilistic interpretation. An iterative scheme that alternates between a foreground modeling module and a region assignment module is proposed in [26], which considers not only the multi-foreground segmentation but also the figure/ground assignment. Background regions can be automatically disregarded with the decomposition framework proposed in [27]. In [28], a scene image co-segmentation method is designed to segment images in the same scene into regions corresponding to their respective classes.

One closely related work is image co-saliency detection [29–36], which aims to detect salient objects in each image as well as common in the image group. Some co-saliency detection models are only suitable for images pairs [30], others can be applied to multiple images [29,31–36]. Co-saliency detection models usually generate single saliency maps first, then generate inter-image saliency maps [29,32–34] or multi-image saliency maps [30,31], and obtain co-saliency maps by integrating the two kinds of maps. In [35], given single image saliency maps, a two-stage query scheme is used to guide co-saliency maps. In [36], exemplar saliency maps are first generated, and then by both local and global recovery, region-level co-saliency maps are obtained from exemplar saliency maps, the region-level co-saliency maps are refined to get the final pixel-level co-saliency maps. As a common characteristic, image co-saliency detection models exploit correspondences between different images to generate co-saliency maps for all images in the image set.

There are some limitations of the image co-segmentation methods we have mentioned above. For the methods which utilize the correspondences among all the images, they treat the detected/segmented objects in all the images as the same; while our observation indicates that the object detection/segmentation results in complicated images are often unsatisfactory, thus unreliable. For the methods that select similar images first, then calculate the correspondences among the selected images, they may achieve satisfactory results for simple images; but not for complicated images. For it could be quite difficult for these images with clutter background to distinguish objects from background, and the correspondence calculation may be useless, even noisy. These limitations have been noticed in [18] and the propagation method from simple images to complicated images has been used, however failure results may also be propagated. In [22], images are also first ranked by their complexities, and the main difference between [22] and the proposed method is the use of simple images. [22] utilizes simple images to select the features and to replace the fixed region similarity measurement for the better region matching, while the proposed method utilizes simple images to extract samples and guide the complicated image segmentation. Besides, the image ranking method is also different from that in [22].

The proposed method pays different regards on simple and complicated images. The simple images are to have better segmentation results which deserve more confidence; while the complicated images, on the contrary, need the aid of simple images. In the proposed method, the complicated images are segmented with the guidance of simple images.

There are two main contributions of the proposed method:

- A new effective method to rank image complexities via their saliency maps.
- An effective framework to utilize simple images to segment complicated images.

The rest of this paper is organized as follows: Section 2 describes the proposed method in detail. Section 3 presents the experimental results and analysis. Finally, Section 4 gives the conclusion.

2. The proposed method

The proposed method includes four steps: *Image Ranking, Simple Image Selection, Samples Extraction* and *Complicated Image Segmentation*, as shown in Fig. 1.

Given an image set$\{I^1, I^2, I^3, \ldots, I^N\}$ composed of N images in total. Firstly, we obtain their single saliency maps $\{S^1, S^2, S^3, \ldots, S^N\}$ by a single saliency detection model [37] and rank the images based on these saliency maps. Secondly, we select the images, in which objects are common and easy to be segmented, as simple images. Thirdly, we co-segment the simple images to obtain their segmentation results and extract the result pieces as samples both positive and negative. Finally, the remaining complicated images are segmented with the guidance of the obtained samples.

To illustrate the whole framework clearly, we introduce some pre-processing work first.

Pre-processing. For each pixel in an image group, we extract its color feature in both RGB and Lab color space, totally in 6 dimensions; and every 4 pixels we extract a dense SIFT descriptor [38] in 128 dimensions. We cluster all the color features into K_c clusters, and all the dense SIFT features into K_s clusters using k-means clustering [39]. Here, K_c and K_s are determined by the two minimum distances, which are set to moderate values, 0.15 and 0.7, for the normalized color features and the normalized SIFT features, respectively. Specifically, during the clustering process, the distance between each cluster pairs is computed, and if there exists a distance shorter than the minimum distance, the closest cluster pair will be merged until no distance between cluster pairs is shorter than the minimum distance. Each pixel has a color cluster label and a dense SIFT cluster label. As shown in Fig. 2, we over-segment each image into three-layer hierarchical regions using gpb-owt-ucm method [40]. For each image, in layer 1, there will be most regions; in layer 2, nearly a half; in layer 3, there will be only about a quarter regions. For an image, there are M regions at most in layer 1 and for a region in each layer, it can be represented by two histograms: color histogram H_c with K_c bins and dense SIFT histogram H_s with K_s bins.

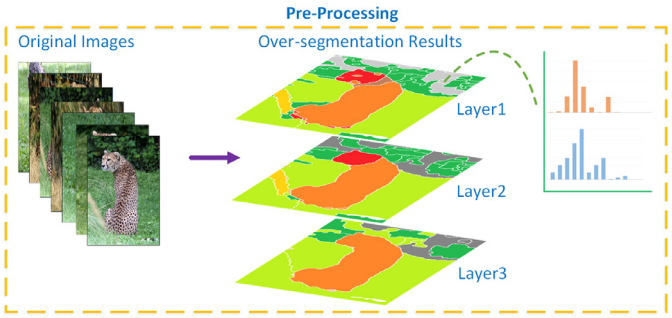


Fig. 2. The pre-processing. One image is over-segmented into hierarchical regions in three layers and is represented by two histograms: color histogram and dense SIFT histogram.

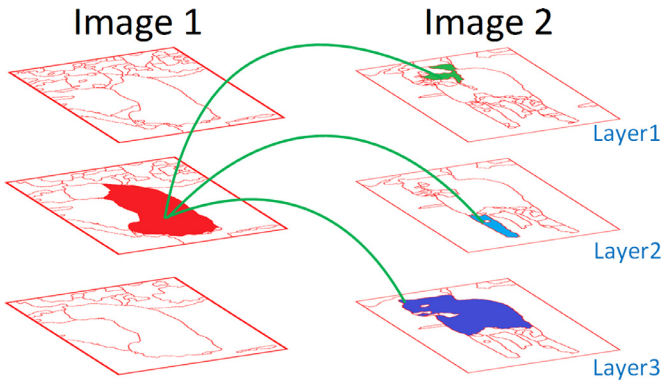


Fig. 3. Region matching. For one region in layer 2 of images 1, its most similar region is searched through all the three layers in image 2.

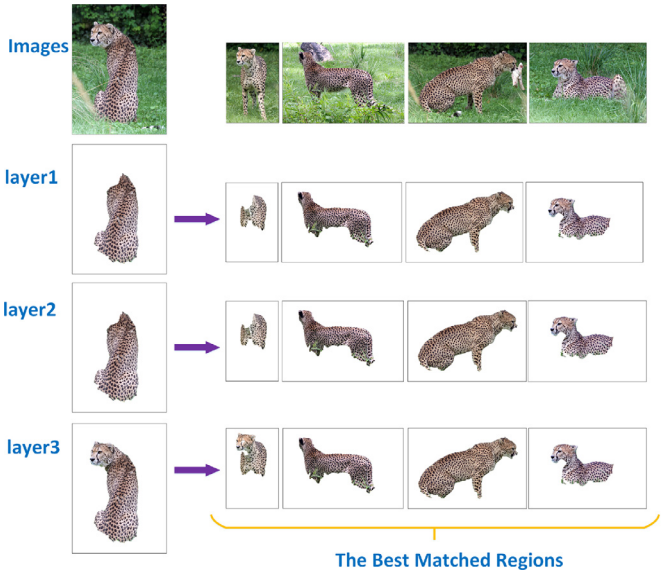


Fig. 4. The region matching results. The 1st row shows original images, the 1st column shows the regions hierarchically extracted from the three layers, and the 2nd–5th column show the corresponding best matched regions in other images.

For region $R^{i,l,n}$, which denotes the ith region in the lth layer of the nth image, its color histogram is denoted as $H_c^{i,l,n}$ and its dense SIFT histogram as $H_s^{i,l,n}$. For two arbitrary regions $R^{i,l,n}$ and $R^{j,k,m}$ in two different images, as shown in Fig. 3, their dissimilarity is calculated as the chi-square distance between their feature histograms as follows:

$$D\left(R^{i,l,n}, R^{j,k,m}\right) = 0.5 \cdot \left(\chi^2\left(H_c^{i,l,n}, H_c^{j,k,m}\right) + \chi^2\left(H_s^{i,l,n}, H_s^{j,k,m}\right)\right) \quad (1)$$

These color and SIFT features are used for the purpose of complementing each other. Fig. 4 shows some experimental region matching results. As we can see, regions in different layers can successfully find their most similarity regions.

2.1. Image ranking

The goal of image ranking is to find the images in which objects are easy to be segmented. We rank images inspired by [41] based

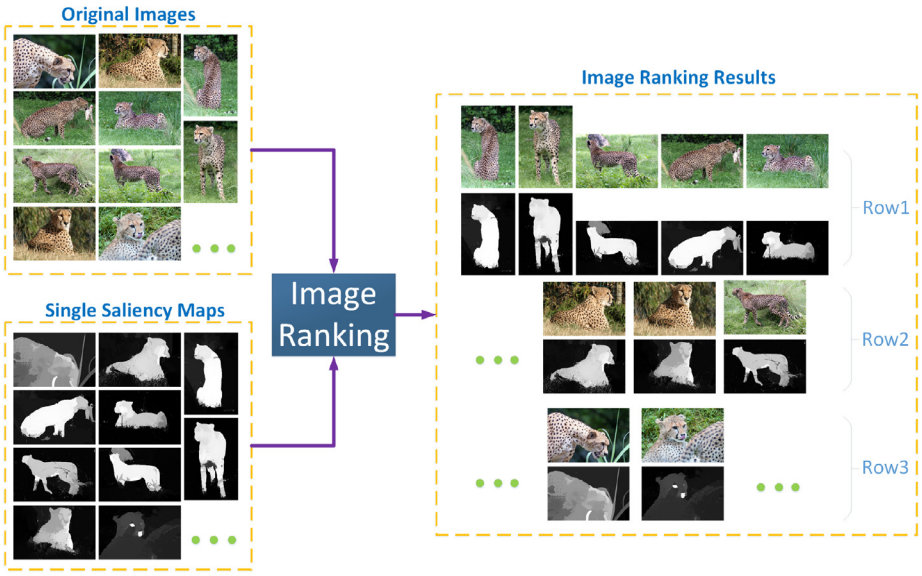


Fig. 5. Image ranking. Images are ranked based on their saliency maps.

on saliency maps. As shown in the right part of Fig. 5, the top ranked images (e.g. the five images in the 1st row) have distinct and uniform saliency maps while the lower ranked ones (e.g. the 2 images in the 3rd row) have bad saliency maps and the objects are not easy to be separated.

The ranking score consists of four elements: *coverage score, compactness score, distribution score* and *contrast score*.

Given a single saliency map S, it is separated into foreground S_f and background S_b by OTSU method [42], i.e.,$\{S_f, S_b\} = OTSU(S)$.

2.1.1. Coverage score

In a high-quality saliency map, the salient region should not cover too much or too little of the whole area. According to the statistical results in [41], we find that the percentage of the S_f occupying the whole area may follow a Gaussian distribution, with the mean value around 0.2 and variance value around 0.2. $P_{cov} = \frac{|S_f|}{|S|}$ and r_{cov} are denoted as the percentage, and coverage score respectively, then

$$r_{cov} = -\frac{(P_{cov} - \mu)^2}{\sigma^2} \tag{2}$$

where $\mu = 0.2, \sigma = 0.2$, they are set based on the statistical results in [41].

2.1.2. Compactness score

A good saliency map should concentrate its salient pixels in a compact region. We search for the corresponding minimum window w_i when the sum of saliency values in w_i occupies a specific percentage of the sum of saliency values in S (denote the percentage as P^i_{com}). Then we compute the average saliency value of w_i (denoted as S^i_{com}). Three windows are searched with $P^i_{com} \in \{0.25, 0.5, 0.75\}, i = 1, 2, 3$. The compactness score r_{com} is computed as a weighted summation as

$$r_{com} = \frac{\sum_i P^i_{com} \cdot S^i_{com}}{\sum_i P^i_{com}} \tag{3}$$

2.1.3. Distribution score

A good saliency map should have two distinct peaks of its value statistics. One for the background regions is close to 0 and the other for the foreground regions is close to 1. To measure the distribution of a saliency map, we first compute two histograms, H_f and H_b for S_f and S_b, respectively. We set the two histograms both with B_d bins. For the b^{th} bin with the histogram value h_b, its weighted histogram value $r_f(b)/r_b(b)$ for foreground/background is calculated by

$$\begin{cases} r_f(b) = h_b \cdot e^{0.5 \cdot (b - B_d)}, b \in H_f \\ r_b(b) = h_b \cdot e^{-0.5 \cdot (b-1)}, b \in H_b \end{cases} \tag{4}$$

where $b = 1, 2, \ldots, B_d$. In our experiments, we set $B_d = 8$. Eq. (4) indicates that in the foreground histogram, one bin which is closer to bin B_d has a higher weight; in the background histogram, one bin which is closer to the 1st bin has a higher weight. The distribution score is calculated as

$$r_{dis} = \sum_{b=1}^{B_d} r_f(b) + \sum_{b=1}^{B_d} r_b(b) \tag{5}$$

A saliency map with distinct peaks in the value statistics can achieve a high distribution score.

2.1.4. Contrast score

A simple image should have a large contrast between its foreground and background regions. For each image, RGB channels are quantized into ${B_c}^3$ bins, with each channel quantized into B_c bins. Then the foreground region S_f and background region S_b are transformed into normalized color histograms hc_f and hc_b, respectively. The contrast score is computed as the chi-square distance between the two histograms

$$r_{con} = \chi^2(hc_f, hc_b) \tag{6}$$

The larger the distance, the higher the contrast score.

2.1.5. Final ranking score

Finally, the ranking score is defined as

$$R = r_{cov} \cdot r_{com} \cdot r_{dis} \cdot r_{con} \tag{7}$$

for the reason that images which can simultaneously obtain four high scores may have good quality. The higher the ranking score, the better quality the images may have.

Fig. 6 shows some ranking results based on different saliency maps. The three kinds of saliency maps are obtained using three state-of-the-art saliency models: ST [37], HS [43] and RBD [44].

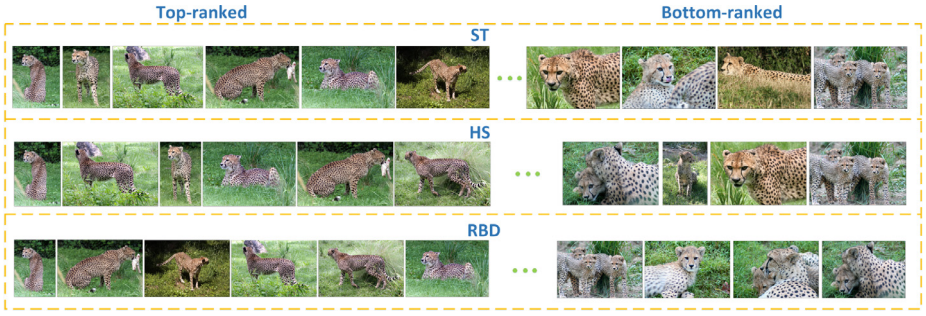


Fig. 6. Image ranking results with different saliency models. The 1st, 2nd and 3rd row shows the ranking results based on saliency maps generated using ST [37], HS [43] and RBD [44], respectively.

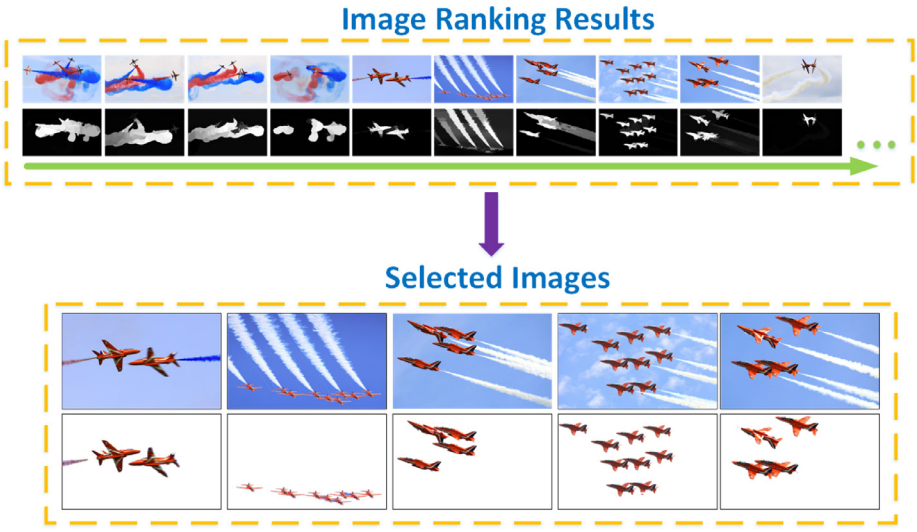


Fig. 7. Simple image selection. There are two boxes. Top box: the original ranked images and their corresponding single saliency maps. Bottom box: the selected simple images and their corresponding co-segmentation results. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

The experimental results[1] show that the ranking results vary with the saliency maps, but in fact there is no essential difference. Although some images will change their ranks, simple images will not be ranked to the bottom as very complicated images, and complicated images will not be ranked to the top as simple images to guide the segmentation of the other images. Therefore, our image ranking method is overall robust to saliency maps generated using different saliency models.

2.2. Simple image selection

In this section, simple image selection aims to select a few images that contain common and easily segmented objects, so their segmentation results can guide the segmentation of the others. Although we have already ranked the images, there may be some noisy images in the top ones that highlight the uncommon regions in the image groups. As shown in Fig. 7, the first ranked 4 images are not selected as simple images for uncommon regions (the red and the blue clouds) are detected.

We use global consistency to guarantee that the segmented objects from the simple images are common in the whole image set.

Global consistency for each region is calculated among the first half of the ranked images. There are two advantages: (1) to reduce the computation cost and (2) the images among the first half are more reliable compared with those in the second half. For region $R^{i,l,n}$, the consistency value is calculated as

$$C^{i,l,n} = 1 - \frac{2}{N}\sum_{m=1}^{N/2} \min_{j,k} D\left(R^{i,l,n}, R^{j,k,m}\right) \tag{8}$$

which indicates to find the most similar regions of $R^{i,l,n}$ in other images, then sum up the minimum distances with normalization to obtain the global consistency. A higher global consistency value means a higher frequency that the region appears over the image group. By combining consistency value computed by Eq. (8) with its single saliency value $S_{si}^{i,l,n}$, we can obtain an initial object likelihood $P^{i,-l,n}$

$$P^{i,-l,n} = S_{si}^{i,l,n} \cdot C^{i,l,n} \tag{9}$$

Since we use hierarchical segmentation method, as Fig. 2 shows, i.e., a region in layer 3 can be split into several regions in layer 2 and a region in layer 2 can be split into several regions in layer 1. Following the [43] the regions in the different layers can be viewed as nodes with parent-child relationship in a tree structure graph. Define $P^{i,l,n}$ as object likelihood variable of node $R^{i,l,n}$, and the set P^n contains all variables in image I^n. Like [43], we construct an energy function which consists of two parts: data term $E_D(R^{i,l,n})$ and hierarchy term $E_S(R^{i,l,n}, R^{j,l+1,n})$

$$E(P^n) = \sum_l \sum_i E_D\left(R^{i,l,n}\right) + \sum_l \sum_{i, R^{i,l,n} \subseteq R^{j,l+1,n}} E_S(R^{i,l,n}, R^{j,l+1,n}) \tag{10}$$

[1] Full image ranking results: https://drive.google.com/open?id=0B9WIl2gJ2teEWnZodUMwS1kyQTA.

Fig. 8. Experimental results. In each image group, from top to bottom, RI: ranked images, ISR: initial segmentation results, and FSR: final segmentation results.

The data term E_D is to gather confidence in each layer, and is defined for every node as follows:

$$E_D\left(R^{i,l,n}\right) = ||P^{i,l,n} - P^{i,-l,n}||_2^2 \tag{11}$$

where $P^{i,-l,n}$ is the initial object likelihood value of region $R^{i,l,n}$ calculated by Eq. (9). The hierarchy term E_S considers the consistency between parent and child nodes, and is defined as follows:

$$E_S\left(R^{i,l,n}, R^{j,l+1,n}\right) = ||P^{i,l,n} - P^{j,l+1,n}||_2^2 \tag{12}$$

By minimizing the energy function Eq. (10), we can obtain the fusion result of initial object likelihood. Since the energy function is a convex function, it can be solved via belief propagation.

The initial co-segmentation result of image I^n can be obtained by separating object from the background in layer 1 of P^n, which is denoted as $P^{top,n}$

$$F_{rc} = OTSU(P^{top,n}) \tag{13}$$

We aim to select N_s simple images by comparing the single segmentation result with the initial co-segmentation result. Denote $F_{si} = OTSU(S_{si})$ as the single segmentation result, then we use their Intersection-over-Union(IoU) as the selection criterion

$$IoU = \frac{|F_{si} \bigcap F_{rc}|}{|F_{si} \bigcup F_{rc}|} \geq \theta \tag{14}$$

From the first image in the ranking queue, if its IoU is larger than or equal to θ, the two segmentation results are considered to be similar and reliable, and the image is selected as a simple image. Repeat this step until we obtain N_s simple images. If there are not enough N_s images that can meet Eq. (14), the images will be selected by their IoU values from top to down, to make the selected simple images as reliable as possible.

2.3. Samples extraction

After selecting simple images, we then segment these images and extract samples from the segmentation results.

For each region $R^{i,l,n}$ in the nth simple image, we first search for its most similar regions in the mth simple image as follows:

$$K(m), J(m) = arg\min_{k,j} D\left(R^{i,l,n}, R^{j,k,m}\right) \tag{15}$$

Thus, the common object likelihood value of region $R^{i,l,n}$ can be calculated as follows:

$$P_{co}^{i,-l,n} = \sum_{m=1}^{N_s} P^{J(m),-K(m),m} \cdot \left(1 - D\left(R^{i,l,n}, R^{J(m),K(m),m}\right)\right) \tag{16}$$

where $P^{J(m),-K(m),m}$ is the initial object likelihood calculated by Eq. (9).

For region $R^{i,l,n}$, we have obtained its common object likelihood value $P_{co}^{i,-l,n}$, then we obtain the optimized common object likelihood $P_{co}^{i,l,n}$ for region $R^{i,l,n}$ by minimizing the tree-structure energy function Eq. (10). Denote P_{co}^{n} as the set that contains all the values in image I^n, and $P_{co}^{top,n}$ as the values in layer 1.

We follow the framework of Grabcut [11] to segment the images with three steps: preliminary labels assignment, GMM estimation and energy minimization.

Preliminary labels assignment. We view the object likelihood values as the prior information for the classification between objects and background. Using $OTSU(P_{co}^{top,n})$, we classify the images into object and background regions. The pixels in object regions are assigned to label 1, and the pixels in background regions are assigned to

label 0. For each pixel p,

$$\mathrm{L(p)} = \begin{cases} 1, p \in object \\ 0, p \in background \end{cases} \tag{17}$$

GMM estimation. We construct two GMMs, one for object regions and the other for background regions. With the component number $K = 5$, we construct the GMMs based on the preliminary object and background regions, respectively, with the following form:

$$\mathbf{\Omega}(\mathbf{x}_n|\mu_i, \mathbf{\Sigma}_i, \omega_i) = \sum_{i=1}^{K} \omega_i N(\mathbf{x}_n|\mu_i, \mathbf{\Sigma}_i) \tag{18}$$

where $\mathbf{x}_n$ is the color vector of pixel n, K is the number of Gaussian component, $N(\mathbf{x}_n|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$ is a Gaussian probability density function, $\boldsymbol{\mu}_i$ is the mean vector of data vectors in the same component, $\mathbf{\Sigma}_i$ is the covariance matrix and $\boldsymbol{\omega}_i$ is the weight of Gaussian component.

Thus for each pixel p, we can obtain its negative log-likelihood value based on the two GMMs, respectively,

$$\begin{cases} l_f(p) = -\log(\mathbf{\Omega}_f(p)) \\ l_b(p) = -\log(\mathbf{\Omega}_b(p)) \end{cases} \tag{19}$$

Energy minimization. The Gibbs energy for segmentation is defined as:

$$E(L) = U(L) + V(L) \tag{20}$$

where $U(L)$ is the data term evaluating the probability of each pixel belonging to object or background,

$$U(L) = \sum_{p} \left[l_f(p) \cdot L(p) + l_b(p) \cdot (1 - L(p)) \right] \tag{21}$$

and $V(L)$ is the smooth term considering the contrast of neighboured pixels defined as:

$$V(L) = \sum_{(p,q)\in C} [p \neq q] \beta * e^{-DC(p,q)} \tag{22}$$

where C is the set of neighborhood pixels, $DC(p, q)$ is the distance between p and q computed using Euclidean distance in the Lab color space.

By minimizing Eq. (20), we can obtain new labels of object and background through GMM estimation and energy minimization iteratively.

After obtaining the co-segmentation results of simple images, as shown in Fig. 1, we extract the samples. The object regions in all layers are extracted as positive samples by assigning the object likelihood as 1. The background regions in all layers are extracted as negative samples by assigning the object likelihood as 0.

2.4. Complicated image segmentation

The remaining complicated images are segmented with the guidance of extracted samples which are more reliable than the complicated images themselves.

There are 6 steps to co-segment a remaining complicated image I^n as follows:

step 1) I^n is over-segmented into regions in three hierarchical layers.

step 2) For region $R^{i,l,n}$ in I^n, distance between $R^{i,l,n}$ and all the samples are computed using Eq. (1) by chi-square distance between their feature histograms. i.e., $R^{i,l,n}$ is compared with all the samples extracted from the N_s simple images.

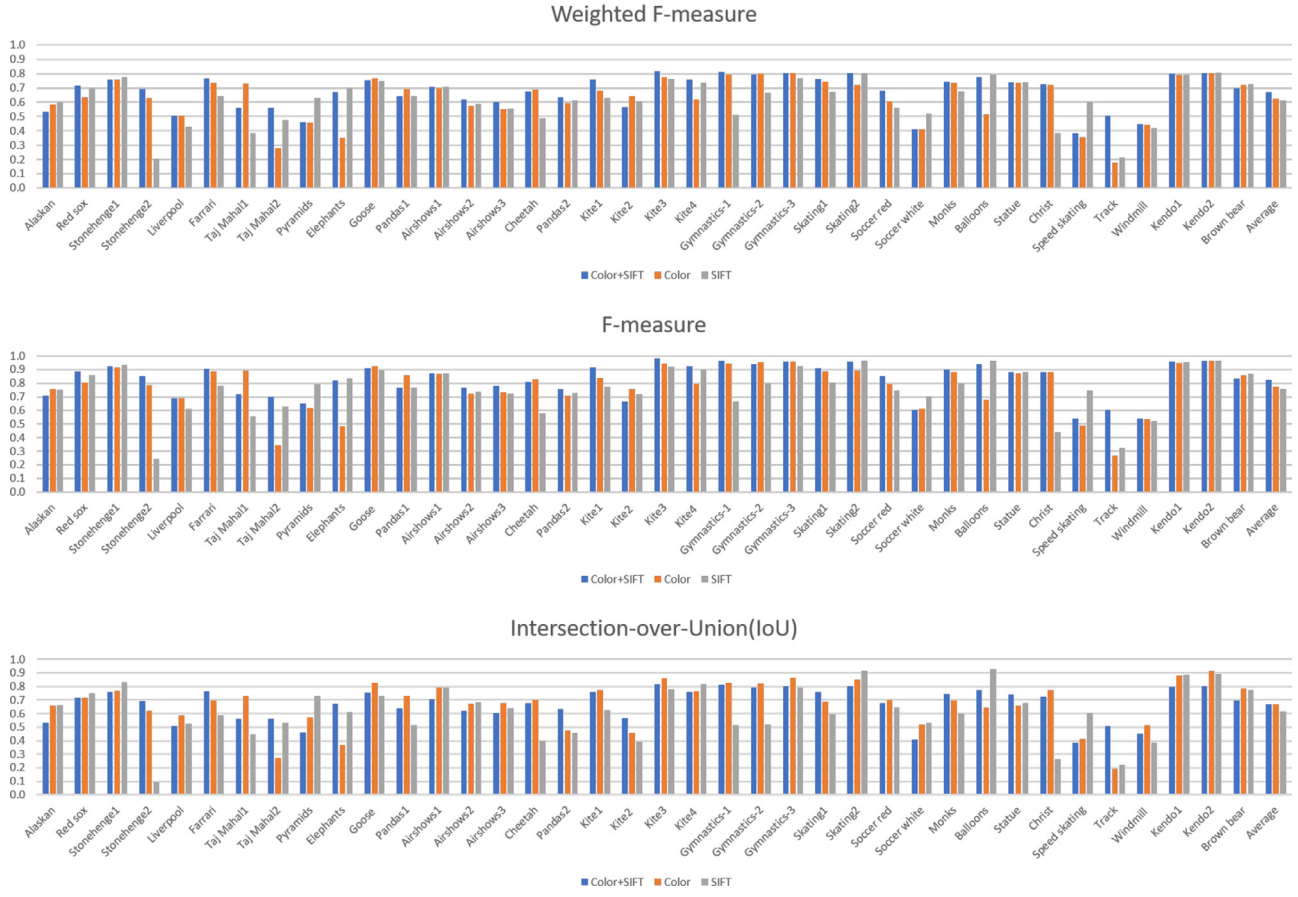


Fig. 9. Performance on iCoseg Dataset with different features.

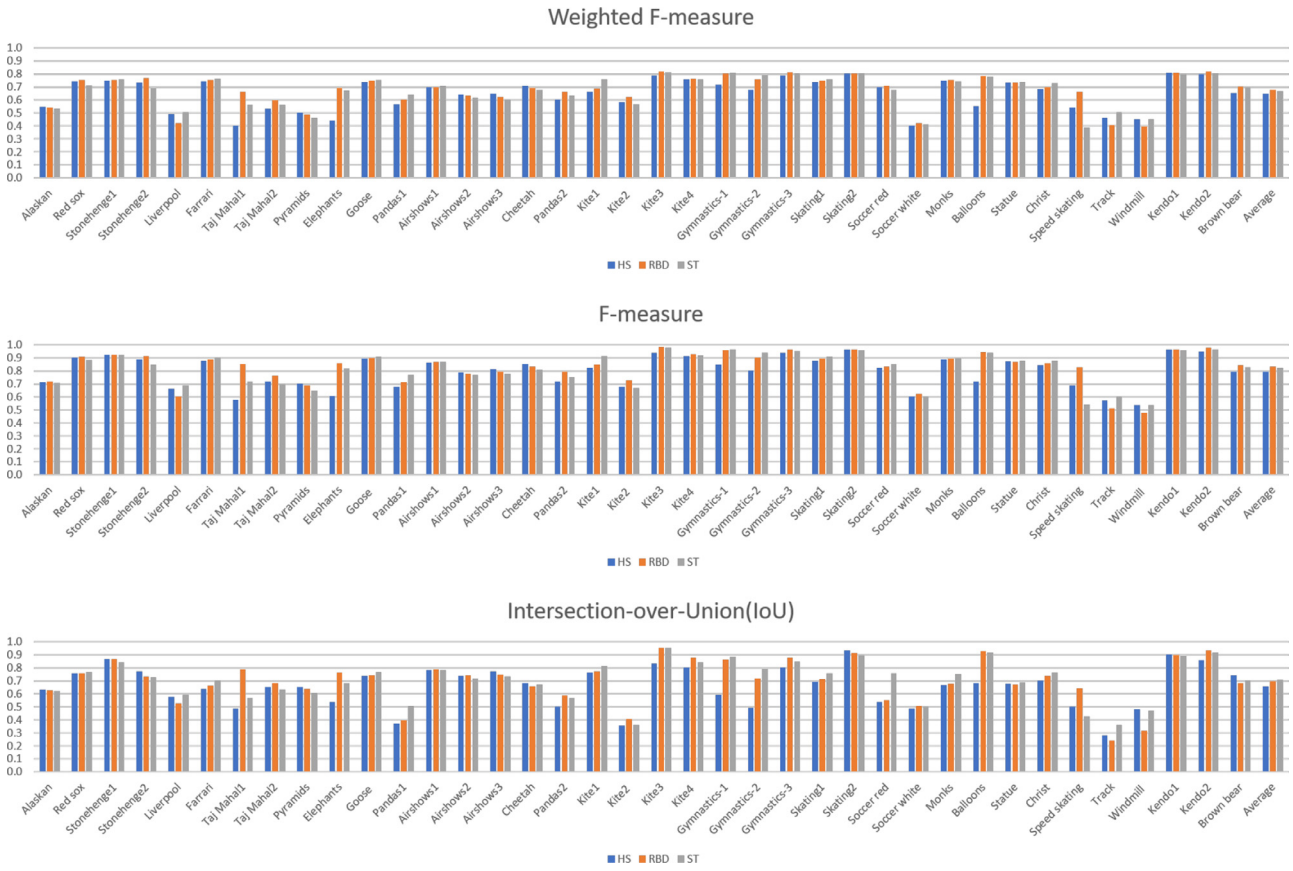


Fig. 10. Performance on iCoseg Dataset with different saliency models.

step 3) A total of N_s most similar samples to $R^{i,l,n}$ are searched using Eq. (15).

step 4) The common object likelihood value of $R^{i,l,n}$ is calculated using Eq. (16). If these samples searched in the previous step are object regions (their object likelihood values are assigned 1), then the common object likelihood value of $R^{i,l,n}$ is probably very high; or contrarily, the samples are background regions, then the common object likelihood value of $R^{i,l,n}$ could be very low.

step 5) Eq. (10) is used to obtain the optimized common object likelihood value of $R^{i,l,n}$.

step 6) Image I^n is segmented through the framework of Grabcut [10]. We use Eq. (17) to get preliminary label assignment, and iteratively adopt object/background GMM estimation and Gibbs energy minimization (from Eqs. (18) to (22)) to obtain the final segmentation results.

Co-segmenting complicated images relying on samples instead of information from the complicated images themselves makes an obvious improvement on the performance, as shown in Figs. 5 and 1. Compare ranking results in Fig. 5 with the complicated image co-segmentation results in Fig. 1: the proposed method can achieve satisfactory segmentation results on the complicated images with low-quality saliency maps, especially for the images in the 3rd row in Fig. 5.

3. Experiments

The experiments are performed on iCoseg dataset [24]. This dataset contains 38 groups with 643 images in total, and each group has 4 to 41 images. We present the experimental results in both subjective and objective ways, and compare the proposed

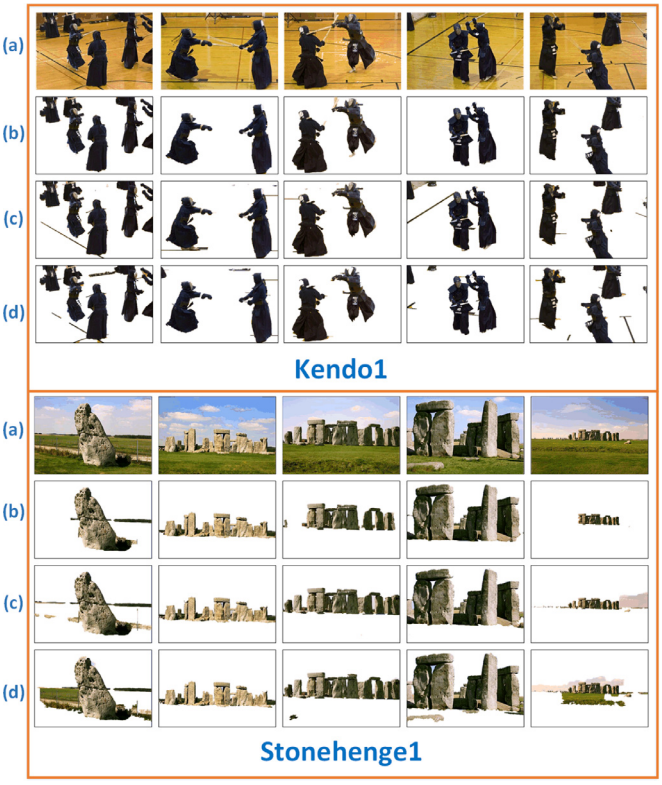


Fig. 11. Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, [21] and [19], respectively.

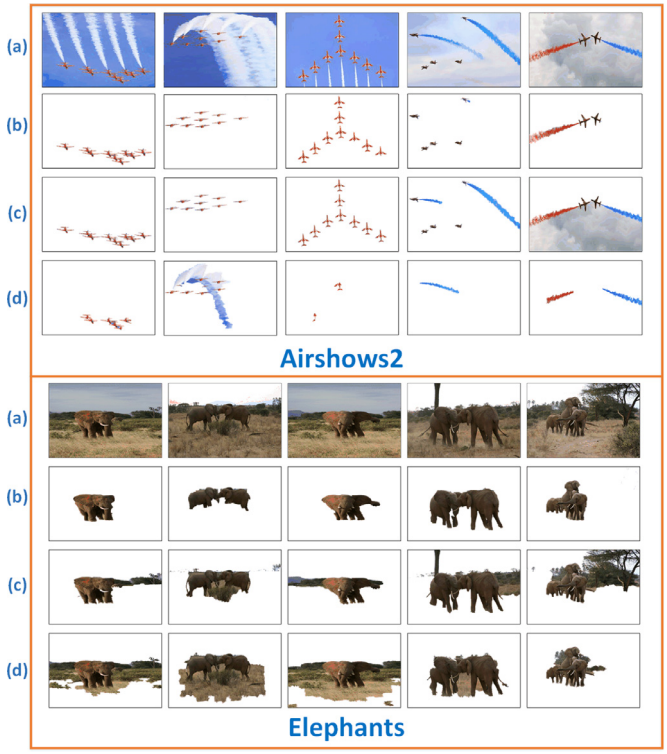


Fig. 12. Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, [21] and [19], respectively.

method with other two state-of-the-art methods [19,21]. In order to have a full set of experimental results, a public code of implementing the reference [19] has been used, therefore the experimental results in our paper are not exactly the same as reported in [19].

To have an overall performance measure, weighted F-measure, Intersection-over-Union (IoU) and F-measure are used in our experiments. The F-measure score is calculated based on four basic quantities: true-positive (TP), true-negative(TN), false-positive (FP) and false-negative (FN) by

$$F_\alpha = \left(1+\alpha^2\right) * \frac{Precision \cdot Recall}{\alpha^2 \cdot Precision + Recall} = \frac{(1+\alpha^2)\cdot TP}{(1+\alpha^2)\cdot TP + \alpha^2 \cdot FN + FP} \quad (23)$$

In our experiments, we value precision more than recall, and as it is suggested in [45], we set $\alpha^2 = 0.3$, the reason for weighting precision more than recall is that recall rate is not as important as precision. For instance, 100% recall can be easily achieved by setting the whole region to foreground. The weighted F-measure is proposed as a more preferable method to evaluate object segmentation results in [46], which replaces the four quantities with weighted quantities and weights errors according to the location and neighborhood.

3.1. Parameter setting

In our experiments, we set the number of simple images $N_s = 5$, the maximum region number $M = 100$, the threshold $\theta = 0.7$ for IoU in simple image selection, and the parameter to control the data term and smooth term in energy minimizing $\beta = 0.2$.

3.2. Results of the proposed method

To more clearly present the guidance effect of simple images, we show the results from the view of simple/complicated images, and the view of initial/final segmentation in Fig. 8, which contains all the selected simple images and the most complicated images. We denote RI as ranked images, ISR as initial segmentation results and FSR as final segmentation results. We can find that the object and background regions can be roughly separated already in the initial segmentation results and the final segmentation results show the refinements. The effective initial segmentation is owed to the correct guidance of the simple images and the effectiveness of region matching.

As we mentioned in *Pre-processing*, we use both color and SIFT features to measure the appearance similarity between regions. These two features can complement each other to some extent. Fig. 4 has showed some results of region matching, and we can also illustrate the matching performance in Fig. 8. In “Cheetah”, the object regions share higher similarity in color, but in “Stonehenge 2”, the objects vary in color due to the change of illumination (e.g. the object in the rightmost image shows different color from the others). In the case of “Stonehenge 2”, the reason why we can segment the object regions is that we also use SIFT feature, and thus the object can be correctly segmented in initial segmentation results. We also compare the results of using both features with the results of using color or SIFT features alone in Fig. 9, and apparently, using both features can achieve the better performance.

To objectively evaluate the sensitivity of the proposed method to saliency maps generated using different saliency models, the IoU values, F-measures and weighted F-measures achieved with the three different saliency models including HS [43], RBD [44] and ST [37] are shown in Fig. 10. From the average scores in Fig. 10, RBD achieves the best both on F-measure and weighted F-measure, ST achieves the best on IoU, while all the three measures with HS are relatively lower compared with RBD and ST. Such a difference with different saliency models is due to the quality of saliency maps. According to the recent benchmark on saliency models [45], RBD and ST are within the top six high-performing models, while the performance of HS is lower than RBD and ST. Nonetheless, the average scores of the three measures among the three saliency models do not show substantial differences. Therefore, we can conclude that the proposed method shows the robustness to different saliency models with similar performance, and the better saliency maps generally lead to the better segmentation results.

For simple image groups with simple backgrounds, as Fig. 11 shows, the proposed method achieves a good performance in the two image groups, and also obtains high scores in the weighted F-measure, IoU and F-measure scores (see “Kendo1” and “Stonehenge1” in Figs. 14–16).

For the complicated image groups which contain many complicated images or multiple objects as shown in Fig. 12, the proposed method can obtain the cleaner object regions, e.g. in “Airshow2”, the planes can be separated with fewer irrelevant regions; in group “Elephants”, few irrelevant background regions are segmented into the objects.

For image groups in which there are not enough real simple images, and the extracted samples are not correct (i.e., some object regions are extracted as negative samples while some background regions are extracted as positive samples), then the guidance of simple images will not lead to a correct cosegmentation of complicated images. For example, as shown in Figs. 8 and 13, in “Speed skating”, background regions (e.g. the advertisement sign) in the simple images are extracted as object regions in the initial segmentation results, and then the background regions (e.g. the advertisement sign) in the complicated images are also segmented as object regions.

Fig. 13. Experimental results. (a) the original images, from (b) to (d), the results of the proposed method, [21] and [19], respectively.

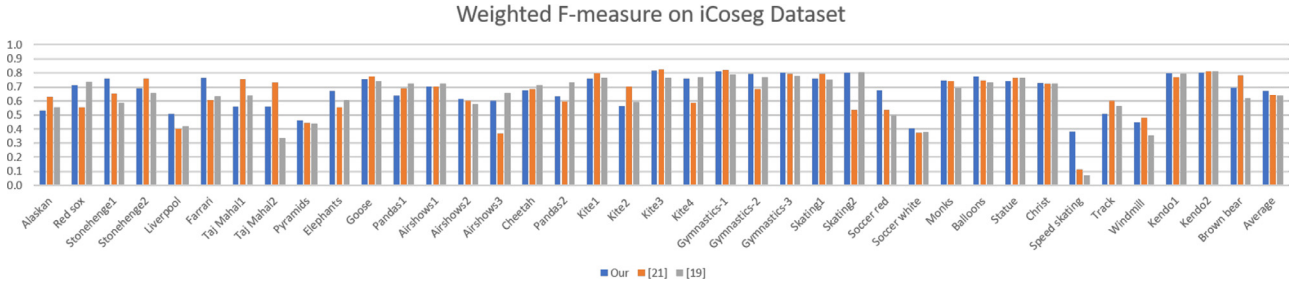


Fig. 14. Weighted F-measure on iCoseg Dataset.

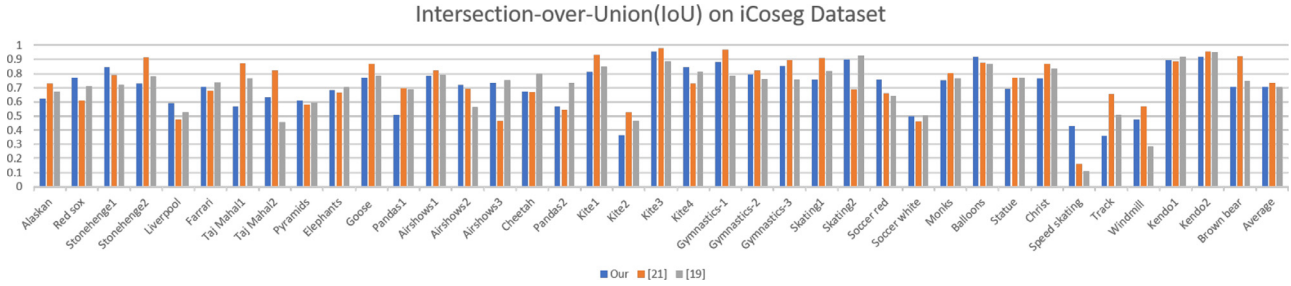


Fig. 15. Intersection-over-Union(IoU) on iCoseg Dataset.

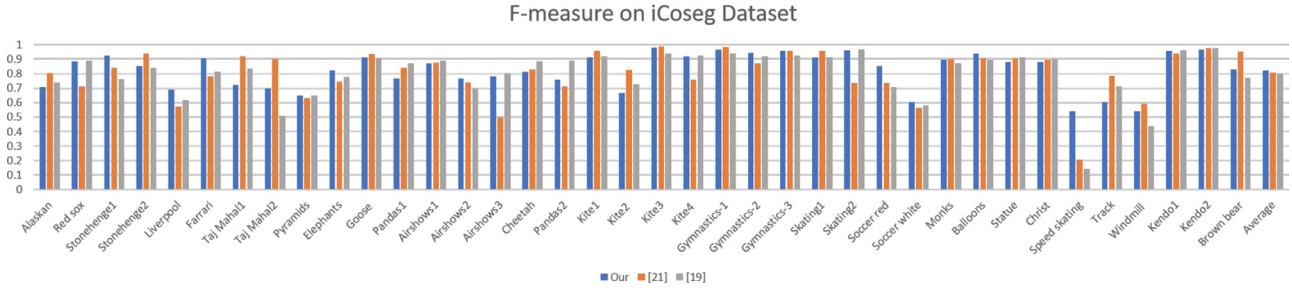


Fig. 16. F-measure on iCoseg Dataset.

3.3. Comparison with other methods

We compare the proposed method with [19,21], and show some subjective results in Figs. 11–13, and objective results in Figs. 14–16. For comparison with [19,21], the saliency model ST [37] is used to generate single saliency maps in the proposed method.

For simple image groups as Fig. 11 shows, the proposed method achieves a good performance on the two image groups with simple backgrounds. In “Kendo1”, the three methods can achieve satisfactory results. In “Stonehenge1”, the results obtained using [21] and [19] contain some irrelevant background regions (e.g. the 1st and 5th column). The results shown in Fig. 11 are in consistency with

Table 1
Overall performance on iCoseg Dataset.

Average Scores	Our	[21]	[19]
Weighted F-measure	0.670	0.645	0.639
F-measure	0.821	0.807	0.800
IoU	0.706	0.734	0.704

Figure 14–16. The three methods can all achieve good segmentation results and high scores on simple image groups.

For complicated image groups, which contain many complicated images, as Fig. 12 shows, the proposed method can obtain the better results than the other two methods. For example, in "Airshow2", the proposed method can segment the planes with few irrelevant regions; the method [21] segments the objects with large irrelevant background regions such as those in the 4th and 5th column; and the method [19] fails to segment the complete objects. In group "Elephants", our results are clear with few irrelevant background regions while the results of the other two methods contain much more irrelevant regions. The results shown in Fig. 12 are in consistency with Figs. 14–16, in which the proposed method achieves the overall higher scores in these groups.

In Fig. 13, none of the three methods can achieve satisfactory results. Nonetheless, the proposed method achieves the relatively better segmentation quality (e.g. the results in the 1st and 4th column) compared with the other two methods, which totally fail to segment the accurate objects from background in this image group. And for the objective evaluation, the proposed method obtains much higher scores on weighted F-measure, IoU and F-measure.

We can observe from the average scores in Table 1 that the proposed method ranks the 1st in terms of F-measure and weighted F-measure, and the 2nd in terms of IoU. In contrast, the method in [21] ranks the 2nd in terms of F-measure and weighted F-measure, and the 1st in terms of IoU. Such an observation is due to the relatively higher precision of the proposed method and the relatively higher recall of the method in [21]. Specifically, for the proposed method and the method in [21], we calculated the average precision as 0.871 and 0.813, respectively, and the average recall as 0.792 and 0.882, respectively. As a comprehensive and overall evaluation using all the three measures, it can be concluded that the proposed method outperforms the other two methods on segmentation performance.

3.4. Discussion

From the experimental results, we can see that utilizing the information of simple images can truly help improve the performance of image co-segmentation, especially for the complicated images; the color feature and SIFT feature can complement each other in region matching, and using them together obtains the better performance than using either feature alone. There are also some limitations of the proposed method. If the objects have a very large variance of appearance or there are not enough really simple images (e.g. "Speed skating" in Figs. 8 and 13), the guidance of simple images will be insufficient, and thus the performance of the proposed method may decrease.

4. Conclusion

In this paper, we propose a novel image co-segmentation method via simple image guidance. An effective image ranking method is introduced and adopted to rank image complexities based on saliency maps. Then samples are extracted by selecting and segmenting simple images. The complicated images are segmented with the guidance of these samples. The motivation of the proposed method is to utilize the information of simple images instead of the information from all the images. From the experimental results, the proposed method has shown its out performance by taking advantages of simple images. We believe that this idea can be further explored and be applied not only in image co-segmentation but also other applications.

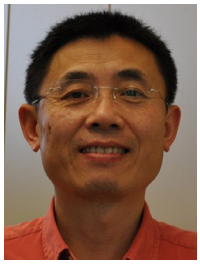
References

[1] G.-H. Liu, J.-Y. Yang, Content-based image retrieval using color difference histogram, Pattern Recognit. 46 (1) (2013) 188–198.
[2] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai, 3-d object retrieval and recognition with hypergraph analysis, IEEE Trans. Image Process. 21 (9) (2012) 4290–4303.
[3] S.-S. Lin, I.-C. Yeh, C.-H. Lin, T.-Y. Lee, Patch-based image warping for content-aware retargeting, IEEE Trans. Multimedia 15 (2) (2013) 359–368.
[4] V. Setlur, T. Lechner, M. Nienhaus, B. Gooch, Retargeting images and video for preserving information saliency, IEEE Comput. Graphics Appl. 27 (5) (2007) 80–88.
[5] S. Zhang, J. Huang, H. Li, D.N. Metaxas, Automatic image annotation and retrieval using group sparsity, IEEE Trans. Syst. Man Cybern. Part B Cybern. 42 (3) (2012) 838–849.
[6] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, S. Ma, Dual cross-media relevance model for image annotation, in: Proceedings of the 15th international conference on Multimedia, ACM, 2007, pp. 605–614.
[7] J. Xue, C. Li, N. Zheng, Proto-object based rate control for JPEG2000: an approach to content-based scalability, IEEE Trans. Image Process. 20 (4) (2011) 1177–1184.
[8] L. Shen, Z. Liu, Z. Zhang, A novel H.264 rate control algorithm with consideration of visual attention, Multimedia Tools Appl. 63 (3) (2013) 709–727.
[9] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.
[10] Y.Y. Boykov, M.-P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in nd images, in: Proceedings of the 8th IEEE International Conference on Computer Vision, vol. 1, IEEE, 2001, pp. 105–112.
[11] C. Rother, V. Kolmogorov, A. Blake, Grabcut: interactive foreground extraction using iterated graph cuts, ACM Trans. Graphics (TOG) 23 (3) (2004) 309–314.
[12] B. Fulkerson, A. Vedaldi, S. Soatto, Class segmentation and object localization with superpixel neighborhoods, in: Proceedings of the 12th International Conference on Computer Vision, IEEE, 2009, pp. 670–677.
[13] C. Rother, T. Minka, A. Blake, V. Kolmogorov, Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs, in: Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, IEEE, 2006, pp. 993–1000.
[14] Y. Li, J. Liu, Z. Li, H. Lu, S. Ma, Object co-segmentation via salient and common regions discovery, Neurocomputing 172 (2016) 225–234.
[15] B.-C. Lin, D.-J. Chen, L.-W. Chang, Unsupervised image co-segmentation based on cooperative game, in: Proceedings of the Computer Vision–ACCV 2014, Springer, 2015, pp. 51–63.
[16] A. Joulin, F. Bach, J. Ponce, Discriminative clustering for image co-segmentation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 1943–1950.
[17] M. Rubinstein, A. Joulin, J. Kopf, C. Liu, Unsupervised joint object discovery and segmentation in internet images, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 1939–1946.
[18] M. Chen, S. Velasco-Forero, I. Tsang, T.-J. Cham, Objects co-segmentation: propagated from simpler images, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 1682–1686.
[19] C. Lee, W.-D. Jang, J.-Y. Sim, C.-S. Kim, Multiple random walkers and their application to image cosegmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3837–3845.
[20] H. Yu, X. Qi, Unsupervised cosegmentation based on superpixel matching and fastgrabcut, in: Proceedings of the International Conference on Multimedia and Expo (ICME), IEEE, 2014, pp. 1–6.
[21] A. Faktor, M. Irani, Co-segmentation by composition, in: Proceedings of the International Conference on Computer Vision (ICCV), IEEE, 2013, pp. 1297–1304.
[22] F. Meng, H. Li, K.N. Ngan, L. Zeng, Q. Wu, Feature adaptive co-segmentation by complexity awareness, IEEE Trans. Image Process. 22 (12) (2013) 4809–4824.
[23] W. Zou, Z. Liu, K. Kpalma, J. Ronsin, Y. Zhao, N. Komodakis, Unsupervised joint salient region detection and object segmentation, IEEE Trans. Image Process. 24 (11) (2015) 3858–3873.
[24] D. Batra, A. Kowdle, D. Parikh, J. Luo, T. Chen, iCoseg: interactive co-segmentation with intelligent scribble guidance, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 3169–3176.
[25] A. Joulin, F. Bach, J. Ponce, Multi-class cosegmentation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 542–549.
[26] G. Kim, E.P. Xing, On multiple foreground cosegmentation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 837–844.
[27] H.-S. Chang, Y.-C. F. Wang, Optimizing the decomposition for multiple foreground cosegmentation, Comput. Vis. Image Underst. 141 (2015) 18–27.
[28] Z. Yuan, T. Lu, P. Shivakumara, A novel topic-level random walk framework for scene image co-segmentation, in: Proceedings of the Computer Vision–ECCV 2014, Springer, 2014, pp. 695–709.

[29] D. Zhang, J. Han, C. Li, J. Wang, Co-saliency detection via looking deep and wide, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2994–3002.
[30] H. Li, K.N. Ngan, A co-saliency model of image pairs, IEEE Trans. Image Process. 20 (12) (2011) 3365–3375.
[31] H. Fu, X. Cao, Z. Tu, Cluster-based co-saliency detection, IEEE Trans. Image Process. 22 (10) (2013) 3766–3778.
[32] S. Du, S. Chen, Detecting co-salient objects in large image sets, Signal Processing Letters, IEEE 22 (2) (2015) 145–148.
[33] Z. Liu, W. Zou, L. Li, L. Shen, O. Le Meur, Co-saliency detection based on hierarchical segmentation, IEEE Signal Process. Lett. 21 (1) (2014) 88–92.
[34] L. Li, Z. Liu, W. Zou, X. Zhang, O. Le Meur, Co-saliency detection based on region-level fusion and pixel-level refinement, in: Proceedings of the International Conference on Multimedia and Expo (ICME), IEEE, 2014, pp. 1–6.
[35] Y. Li, K. Fu, Z. Liu, J. Yang, Efficient saliency-model-guided visual co-saliency detection, IEEE Signal Process. Lett. 22 (5) (2015) 588–592.
[36] L. Ye, Z. Liu, J. Li, W.-L. Zhao, L. Shen, Co-saliency detection via co-salient object discovery and recovery, IEEE Signal Process. Lett. 22 (11) (2015) 2073–2077.
[37] Z. Liu, W. Zou, O. Le Meur, Saliency tree: A novel saliency detection framework, IEEE Trans. Image Process. 23 (5) (2014) 1937–1952.
[38] A. Vedaldi, B. Fulkerson, Vlfeat: An open and portable library of computer vision algorithms, in: Proceedings of the International Conference on Multimedia, ACM, 2010, pp. 1469–1472.
[39] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137.
[40] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, Contour detection and hierarchical image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 898–916.
[41] L. Mai, F. Liu, Comparing salient object detection results without ground truth, in: Proceedings of the Computer Vision–ECCV, Springer, 2014, pp. 76–91.
[42] N. Otsu, A threshold selection method from gray-level histograms, Automatica 11 (285–296) (1975) 23–27.
[43] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 1155–1162.
[44] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2814–2821.
[45] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5706–5722.
[46] R. Margolin, L. Zelnik-Manor, A. Tal, How to evaluate foreground maps? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 248–255.

Lina Li received the B.E. degree from Yangzhou University, Yangzhou China, in 2011. She is currently pursuing the Ph.D. degree at Shanghai University and the University of Technology, Sydney. Her current research interests include co-saliency detection, image co-segmentation and video co-segmentation.

Zhi Liu received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, in 1999, 2002, and 2005, respectively. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 150 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision and multimedia communication. He was a TPC member in VCIP 2016, ICME 2014, WIAMIS 2013, IWVP 2011, PCM 2010, ISPACS 2010, etc. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an area editor of *Signal Processing: Image Communication* and served as a guest editor for the special issue on Recent Advances in Saliency Models, Applications and Evaluations in *Signal Processing: Image Communication*. He is a senior member of IEEE.



Jian Zhang received the B.S. degree from East China Normal University, Shanghai, China, in 1982; the M.S. degree in computer science from Flinders University, Adelaide, Australia, in 1994; and the Ph.D. degree in electrical engineering from the University of New South Wales (UNSW), Sydney, Australia, in 1999.

From 1997 to 2003, he was with the Visual Information Processing Laboratory, Motorola Labs, Sydney, as a Senior Research Engineer, and later became a Principal Research Engineer and a Foundation Manager with the Visual Communications Research Team. From 2004 to July 2011, he was a Principal Researcher and a Project Leader with National ICT Australia, Sydney, and a Conjoint Associate Professor with the School of Computer Science and Engineering, UNSW. He is currently an Associate Professor at the Global Big Data Technologies Centre, School of Electrical & Data Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney. He is the author or co-author of more than 130 paper publications, book chapters, and eight granted US and China patents. His current research interests include social multimedia signal processing, image/video based scene understanding, large scale data analytics for industry application, and intelligent video surveillance systems.

Dr. Zhang was the General Co-Chair of the International Conference on Multimedia and Expo in 2012 and Technical Program Co-Chair of IEEE Visual Communications and Image Processing 2014. He is a senior member of IEEE and Associated Editors for the IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2006-2015) and the EURASIP Journal on Image and Video Processing.