Coarse-to-fine Image Co-segmentation with Intra and Inter Rank Constraints

Lianli Gao, Jingkuan Song, Dongxiang Zhang, Heng Tao Shen*
Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
{lianli.gao, zhangdo}@uestc.edu.cn, jingkuan.song@gmail.com, shenhengtao@hotmail.com

Abstract
Image co-segmentation is the problem of automatically discovering the common objects co-occurring in a set of relevant images and segmenting them as foreground simultaneously. Although a bunch of approaches have been proposed to address this problem, many of them still suffer from certain limitations, e.g., supervised feature learning and complex models, which hinder their capability in the real-world scenarios. To alleviate these limitations, we propose a novel coarse-to-fine co-segmentation (CFC) framework, which utilizes the coarse foreground and background proposals to learn a robust similarity measure of the features in an unsupervised way, and then devises a simple objective function based on the definition of image co-segmentation. Specifically, we first generate superpixels for all the images and extract their features. Instead of using existing distance metrics, we utilize object proposal methods to generate coarse foreground and background to learn a similarity measure of superpixels to construct a robust feature similarity graph. Then we design an intuitive objective function to learn a segmentation similarity graph which should be consistent with feature similarity graph and also be able to co-segment the superpixels in the images into either foreground and background. This objective function can be further reformulated as a graph learning problem with intra and inter rank constraints. Experiments on two commonly used image datasets (iCoseg and MSRC) demonstrate that CFC outperforms other state-of-the-art methods. Notably, this performance is achieved by using only HSV feature.

1 Introduction
Co-segmentation, i.e., jointly segmenting the common objects from a collection of similar images, has attracted an increasing attention recently in computer vision community.

*Corresponding author: Lianli Gao, Heng Tao Shen

[1] Vicente, J. A., 2015 and discriminative clustering [Joulin et al., 2010]. Compared with single image segmentation [Gao et al., 2016], co-segmentation has the advantage of utilizing the very weak prior that the set of images contain the same objects to improve the segmentation of individual images. At the meanwhile, co-segmentation is a challenging problem, which faces several major issues: 1) The classical segmentation methods are designed for a single image while co-segmentation deals with multiple images. How to design a co-segmentation model and minimize the model is crucial for image co-segmentation. 2) The common object extraction depends on the foreground similarity measurement. But the foreground usually varies in color, shape and scale, which makes the foreground similarity measurement difficult. Moreover, the model minimization is highly related to the selection of similarity measurement. Therefore, the foreground similarities measurement is another important issue.

A straight-forward way for co-segmentation is to extend classical single image based segmentation models. In general, the extended models can be represented as

$$\ell = \ell_s + \ell_g$$

where $\ell_s$ is the single image segmentation term, which guarantees the smoothness and the distinction between foreground and background in each image, and $\ell_g$ is the co-segmentation term, which focuses on evaluating the consistency between the foregrounds among the images.

Many classical single image segmentation models can be used to form $\ell_s$, e.g. MRF segmentation models [Lee et al., 2015] and discriminative clustering [Joulin et al., 2010]. The co-segmentation term $\ell_g$ is used to evaluate the multiple foreground consistency, which is introduced to guarantee the common object segmentation. Due to the variance of the foreground, the similarity measure using visual features is not accurate enough.

On the other hand, segmentation proposal selection based approaches [Vicente et al., 2011; Meng et al., 2012; 2013] have recently drawn more attention. These methods aim at select the real common objects from the segmentation proposals, and the assumption that the common targets should be real-world objects makes co-segmentation much more efficient. Compared with co-segmenting images using pix-
els, segmentation proposal selection based approaches have their advantages. First, instead of starting from scratch, they incorporate either class-specific or class-independent object proposal methods to generate a set of candidates, and then look for a pool of regions that have high probability to cover the objects by some of the proposals. The proposals are usually more robust, and they contain rich information than pixels. Therefore, they usually achieve better accuracy than methods based on pixels. Secondly, they are more efficient compared with pixel-based methods, because the number of proposals is usually much smaller than the number of pixels in images. However, their disadvantages are obvious as well. They usually reply heavily on the performance of object proposals. Also, due to the variety of the common objects, it usually requires manually selecting features [Meng et al., 2012] or supervised feature learning performed beforehand [Vicente et al., 2011; Meng et al., 2013].

To address the above issues, in this work, we propose a novel coarse-to-fine image co-segmentation framework, which introduces graph-based image co-segmentation using superpixels based on intra and inter rank constraints. To take advantage of proposal, we utilize the coarse foreground and background proposals to learn a robust similarity measure of the superpixels. It is worthwhile to highlight the following aspects of our method here: 1) We propose a novel coarse-to-fine image co-segmentation framework. To take advantage of object proposals but avoid heavily depending on the proposal quality, we utilize multiple image proposal strategies and generate a set of coarse foreground and background. These incomplete but with few noise coarse segments are further used to learn a similarity measurement of the raw features to obtain a fine image co-segmentation. 2) We design an intuitive graph-based objective function for image co-segmentation based on its definition. Taking the feature similarity graph as input, we aim to learn a segmentation similarity graph which should be consistent with feature similarity graph and also be able to co-segment the superpixels in images into either foreground or background. We devise an efficient solution for this objective function. 3) Experiments for image co-segmentation on real-world datasets demonstrate the superiority of our method over existing methods. Notably, this performance is achieved using HSV feature only.

2 Our Approach

In this section, we introduce our method which consists of two phases (see Fig. 1). Firstly, we further introduce how to construct an original similarity graph using the coarse foreground and background proposals. Then, a graph-based image co-segmentation algorithm is proposed based on the original similarity graph of the superpixels.

We first introduce the notations which will be used in the rest of the paper. Given a set of M images $I = \{I_1, I_2, ..., I_M\}$ for an object class, we first generate a set of $N$ oversegmenting superpixels $V = \{v_1, v_2, ..., v_N\}$ as a preprocessing, and then extract features $X = \{x_1, x_2, ..., x_N\}$ from these N superpixels. A superpixel is an image segment consisting of pixels that have similar visual characteristics.

The goal is to learn the similarity matrix $S$ between each superpixel based on the features $X$, and all the superpixels are segmented to either foreground or background.

2.1 Coarse Foreground and Background Proposal for Original Similarity Graph Construction

Instead of segmenting common objects from an object class from scratch, we start by generating a set of coarse foreground and background proposals. The motivation is that if we can generate pure foreground and background with high precision but low recall, we can use these regions as coarse proposals to construct an accurate original similarity graph. Inspired by recent progress in object proposal and salient object detection, we find that those two techniques have potential to propose incomplete by with few noise objects in an image. This is because salient object detection (e.g., PCA Saliency [Margolin et al., 2013]) is originally a task of predicting the eye-fixations on images, and recently has been extended to detecting regions with salient attention-grabbing objects, while object proposal (e.g., MCG [Arbeláez et al., 2014]) aims to generate a set of regions to delineate candidate objects in an image. However, current generated saliency maps are typically blurry, especially near the boundary of salient objects, while object proposal methods can provide accurate object boundaries but it is still difficult to obtain complete objects by merging together sets of regions, thus they are complementary to each other.

Therefore, in this study we take the advantages of both salient object detect [Margolin et al., 2013] and object proposal [Arbeláez et al., 2014] to produce coarse foreground and background, which respectively contain incomplete but relatively pure foreground and background, to refine our similarity graph. Here, we introduce the construction details. Given an image $I_g \in \mathbb{R}^{w \times h}$, where $w$ and $h$ are the width and length.

Saliency map and candidate mask generation. At beginning, PCA Saliency [Margolin et al., 2013] takes $I_g$ as input and produce a saliency map $s_{map} \in \mathbb{R}^{w \times h}$, while MCG [Arbeláez et al., 2014] takes $I$ as input and produces a ranked list of object candidates. Here, we merge the top-5 ranked object candidates to output a candidate mask $c_{mask} \in \mathbb{R}^{w \times h}$. If $c_{mask}_{i,j} = 1$, it stands that the pixel $I_g(i,j)$ is an object pixel otherwise it is a background pixel.

Refine candidate with UCMs. After we obtained the initialized $c_{Mask}$, we apply ucms to refine $c_{Mask}$. Firstly, we extract superpixels with $ucm \leq 0.5$ and then remove the superpixels containing border pixels, except those superpixels meets one of the following conditions: having less than $\frac{1}{2}(w + h)$ border superpixels; 2) only containing left or right borderer pixels; and 3) the number of top/bottom borderer pixels is less than $\frac{3}{4}$ of the corresponding boarder length. Next, we project these superpixels to $ucm \leq 0.18$ and then merge with $c_{Mask}$ to obtain a new $c_{Mask}$. Note that, all the parameter values are empirically set as above. Next, we move to coarse foreground and background generation.

Coarse background generation. At first, we compute a threshold $th_b$ by averaging the values of $s_{Map}$ that are lower than the mean of $s_{Map}$. Next, comparing each element of $s_{map}$ with $th_b$ to initialize the coarse background mask.
Coarse foreground generation. First, we look for elements of $s_{map}$ that are larger than $s_{Map}$ mean and then set them as 1, otherwise set as 0, to form an initial coarse foreground mask $f_{mask}$. Next, we joint $f_{mask}$ and $c_{mask}$ to obtain the coarse foreground mask $f_{mask}$. If the element of $f_{mask}$ is 1, the corresponding pixel is pure foreground.

2.2 Image Co-segmentation with Intra and Inter Rank Constraints

Instead of following the conventional pipeline for image co-segmentation, in our framework (see Fig. 1), we propose a novel perspective to solve the graph-based image co-segmentation problem. Based on the definition of image co-segmentation, all the images have two segments (foreground and background), and each image also has two segments. Conventional methods usually construct a similarity graph $A$ based on the features of superpixels, and all the superpixels are connected as just one connected component. Ideally, if the superpixels of all the images (and also for each image) are connected as exactly 2 components, the superpixels are naturally divided into 2 segmentations and this segmentation is the optimal solution for different criteria, e.g., normalized cut [Shi and Malik, 2000]. Then we can get our objective function:

$$\min_S \|S - A\|_F^2$$

$$s_{ij} \geq 0, s_1 = 1,$$

$$S\text{ has 2 connected components},$$

$$S_m\text{ has 2 connected components}$$

(2)

where $S$ is the segmentation similarity graph to be learned, and $A \in \mathbb{R}^{N \times N}$ is the original similarity graph, $S_m$ is a block of $S$ and it is the segmentation similarity matrix of image $I_m$. The object function aims to learn a $S$ which is similar to $A$. To avoid the case that some rows of $S$ are all zeros, we further constrain the $S$ such that the sum of each row of $S$ is one. Under these constraints, we learn a $S$ that best approximates the initial affinity matrix $A$. The last two constraints require that all the images have two segments (foreground and background), and each image also has two segments, which is the definition of image co-segmentation.

The problem in (2) is not easy to solve because of the constraints. To tackle this, we start from the following theorem. For a nonnegative similarity matrix $S$, there is a Laplacian matrix $L$ associated with it. According to the definition of Laplacian matrix, suppose each superpixel $x_i$ is assigned a random value as $f_i \in \mathbb{R}^{1 \times K}$, then $L$ can be calculated as:

$$\sum_{ij} \|f_i - f_j\|^2_s s_{ij} = 2 \text{tr} (F^T L F)$$

(3)

where $F \in \mathbb{R}^{N \times K}$ with the $i$-th row formed by $f_i$, $L = D - \frac{S + S^T}{2}$ is called the Laplacian matrix in graph theory, the degree matrix $D \in \mathbb{R}^{N \times N}$ is defined as a diagonal matrix where the $i$-th diagonal element is $\sum_j (s_{ij} + s_{ji}) / 2$. The Laplacian matrix $L$ has the following property.

Theorem 1 The number $K$ of the eigenvalue 0 of the Laplacian matrix $L$ is equal to the number of connected components in the graph with the similarity matrix $S$ if $S$ is nonnegative.

Theorem 1 indicates that if $\text{rank}(L) = N - 2$, then the superpixels have 2 connected components based on $S$. Thus, (2) can be reformulated as:

$$\min_S \|S - A\|_F^2$$

$$s_{ij} \geq 0, s_1 = 1,$$

$$S\text{ has 2 connected components},$$

$$S_m\text{ has 2 connected components}$$

st. 

$$\text{rank}(L) = N - 2$$

$$\text{rank}(L_m) = N - 2$$

(4)

where $L$ is the laplacian matrix of $S$, and $L_m$ is the laplacian matrix of $S_m$. It is still difficult to solve the problem (4). Because $L = D - \frac{(S^T + S)}{2}$ and $D$ also depends on $S$. The constraint $\text{rank}(L) = N - 2$, $\text{rank}(L_m) = N - 2$ are not easy to tackle. Suppose $e_i$ is the $i$-th smallest eigenvalue of $L$ and $(e_m)_i$ is the $i$-th smallest eigenvalue of $L_m$, we know $e_i, (e_m)_i \geq 0$ since $L$ and $L_m$ are positive semi-definite. It
can be seen that the problem (4) is equivalent to the following problem for a large enough value of $\alpha$ and $\beta$:

$$\min_{s} \|S - A\|_F^2 + 2\alpha \sum_{i=1}^{2} e_i + 2\beta \sum_{i=1}^{2} (e_m)_i$$

s.t. $s_{ij} \geq 0, s_{11} = 1$

(5)

When $\alpha$ and $\beta$ are set to a large enough value, $\sum_{i=1}^{2} e_i$ and $\sum_{i=1}^{2} (e_m)_i$ will be imposed to be close to 0, which results in $\text{rank}(L) = N - 2$ and $\text{rank}(S_m) = N - 2$.

According to the Ky Fan’s Theorem [Fan, 1949], we have:

$$\sum_{i=1}^{2} e_i = \min_{F \in \mathbb{R}^{N \times 2}} \text{tr} (F^T LF)$$

(6)

Therefore, (5) is further equivalent to:

$$\min_{s,F,F_m} \|S - A\|_F^2 + 2\alpha \sum_{m=1}^{M} \text{tr} (F_m^T L_m F_m) + 2\beta \sum_{m=1}^{M} \text{tr} (F_m^T L_m F_m)$$

s.t. $s_{ij} \geq 0, s_{11} = 1$

(7)

By fixing $S$, we can obtain $F$ and $F_m$ by optimizing Eq.7. This is equivalent to optimize the following objective function:

$$\min_{F,F_m} \|S - A\|_F^2 + \alpha \text{tr} (F^T LF) + \beta \sum_{m=1}^{M} F_m^T L_m F_m$$

s.t. $F^T F = I_N, F_m^T F_m = I_{N_m}$

which can be further reformulated as:

$$\min_{F} \text{tr} \left( F^T \left( \alpha L + \beta \sum_{m=1}^{M} U_m L_m U_m^T \right) F \right)$$

s.t. $F^T F = I_N, F_m^T F_m = I_{N_m}$

(8)

This problem can be solved by obtaining the two smallest eigenvector of $(\alpha L + \beta \sum_{m=1}^{M} U_m L_m U_m^T)$, and $F_m = U^T F$.

**Update $S$**

By fixing $F$ and $F_m$, we can obtain $S$ by optimizing Eq.7. It is equivalent to optimize the following objective function:

$$\min_{s,S} \|S - A\|_F^2 + \alpha \sum_{i,j=1}^{N} f_{ij}^2 s_{ij} + \beta \sum_{m=1}^{M} \sum_{p,q=1}^{N_m} (f_{m,p} - (f_m)_{pq})^2 (s_m)_{pq}$$

s.t. $s_{ij} \geq 0, s_{11} = 1$

(10)

where $B_m = U_m V_m U_m^T$ and $(v_m)_{pq} = \left\| (f_m)_{pq} - (f_m)_p \right\|^2$.

Then (10) becomes:

$$\min_{s} \|S - A\|_F^2 + \alpha \sum_{i,j=1}^{N} \left( \alpha \|f_{ij} - f_{ij}\|^2 + \beta \sum_{m=1}^{M} (b_{ij,m})_s \right)$$

s.t. $s_{ij} \geq 0, s_{11} = 1$

(12)

**Iterative Optimization**

We propose an iterative method to minimize the above objective function in Eq.7. We first define the relationships between $S$ to $S_m$, and $F_m$ to $F$. Recall that $S$ is the original similarity graph of all images, and $S_m$ is the similarity graph of the $m$-th image. Therefore, $S_m = S(\text{id}_{s_m} : \text{id}_{m}, \text{id}_{m} : \text{id}_{m})$, where $\text{id}_{s_m}$ and $\text{id}_{m}$ are the index for the starting and ending position of superpixels in $M$ images set. Formally, we define a selection matrix $U \in \mathbb{R}^{N \times N_m}$ so that $S_m = U_m^T S U_m$, where $U$ is a matrix with all zeros, but the $\text{id}_{s_m}$ to $\text{id}_{m}$ rows are filled by an identity matrix $I_{N_m}$. Similarly, we can get $F_m = U_m^T F$.

By taking the original similarity matrix $A$ as input, in each iteration, our algorithm first updates $F$ and $F_m$ given $S$, and then updates $S$ by fixing $F$ and $F_m$. These steps are described below.

**Update $F$ and $F_m$**

By fixing $S$, we can obtain $F$ and $F_m$ by optimizing Eq.7. This is equivalent to optimize the following objective function:

$$\min_{F,F_m} \text{tr} \left( F^T LF \right) + \beta \sum_{m=1}^{M} F_m^T L_m F_m$$

s.t. $F^T F = I_N, F_m^T F_m = I_{N_m}$

(13)

which can be further reformulated as:

$$\min_{F} \text{tr} \left( F^T \left( \alpha L + \beta \sum_{m=1}^{M} U_m L_m U_m^T \right) F \right)$$

s.t. $F^T F = I_N, F_m^T F_m = I_{N_m}$

(14)

**Update $S$**

By fixing $F$ and $F_m$, we can obtain $S$ by optimizing Eq.7. It is equivalent to optimize the following objective function:

$$\min_{s,S} \|S - A\|_F^2 + \alpha \sum_{i,j=1}^{N} f_{ij}^2 s_{ij} + \beta \sum_{m=1}^{M} \sum_{p,q=1}^{N_m} (f_{m,p} - (f_m)_{pq})^2 (s_m)_{pq}$$

s.t. $s_{ij} \geq 0, s_{11} = 1$

(10)

where $B_m = U_m V_m U_m^T$ and $(v_m)_{pq} = \left\| (f_m)_{pq} - (f_m)_p \right\|^2$.

Then (10) becomes:

$$\min_{s} \|S - A\|_F^2 + \alpha \sum_{i,j=1}^{N} \left( \alpha \|f_{ij} - f_{ij}\|^2 + \beta \sum_{m=1}^{M} (b_{ij,m})_s \right)$$

s.t. $s_{ij} \geq 0, s_{11} = 1$

(12)

Denote $g_{i,j} = \alpha \|f_{i,j} - f_{i,j}\|^2 + \beta \sum_{m=1}^{M} (b_{ij,m})$, and (12) can be reformulated as:

$$\min_{s} \sum_{i,j=1}^{N} \left( s_{ij} - a_{ij} \right)^2 + \sum_{i,j=1}^{N} g_{ij} s_{ij}$$

s.t. $s_{ij} \geq 0, s_{11} = 1$

(13)

The problem in Eq.13 is simple [Nie et al., 2016] and the critical step of the projected gradient method is to solve the following proximal problem:

$$\min_{x > 0} \frac{1}{2} \|x - c\|^2$$

This proximal problem can be solved by an efficient iterative algorithm [Nie et al., 2016]. Then each $s_i$ can be efficiently solved, and we can get the updated graph $S$.

We update $F$ and $S$ iteratively until the objective function Eq.7 converges.

**3 Experiments**

We evaluate our algorithm on the task of image cosegmentation. Firstly, we compare our results with state-of-the-art algorithms on two standard datasets. Then, we study the influence of different features for our algorithm.

**3.1 Experimental Settings**

**Datasets**

We evaluated the proposed algorithm on two public benchmark datasets: the iCoseg dataset [Batra et al., 2010] and the MSRC dataset [Winn et al., 2005]. iCoseg dataset contains 38 image classes of totally 643 images with manually labeled pixel-wise ground-truth masks. We follow [Lee et al., 2015] to use a subset of 16 object classes. The later includes 591 pixelwise labeled images of 23 object classes. We select the same classes as reported by [Joulin et al., 2010;
Lee et al., 2015]. The complex background of the MSRC dataset makes it more challenging, and it involves considerable changes in viewpoints and illumination.

Following [Lee et al., 2015], we evaluate accuracy, precision and recall for the segmentation. A segmentation accuracy is defined as the percentage of correctly labeled pixels, while precision and recall are defined as the percentage of correctly labeled foreground.

**Features**

We utilize two kinds of image features, i.e., the shallow features and the deep features, to capture different characteristics of each superpixel. Specifically, we used HSV feature in our work, and we further test another three kinds of shallow features, including LAB, texture, and dense SIFT descriptors. Software from [van de Sande et al., 2011] is used, and the dimensions for each feature is 75-D for HSV, 75-D for LAB, 240-D for texture, and 1024-D for SIFT.

For deep feature, we employ the CNN-S [Chatfield et al., 2014] model which is pre-trained on the ImageNet dataset. Firstly, we feed each image into the pre-trained CNN and get the responses from the last convolutional layer, which consists of 512 feature maps with size of 17×17. Next, we resize the feature maps to the size of the original image, and then use a max pooling operation on each superpixel to generate a 512-D CNN feature vector.

We run our algorithm using single image (IS) and co-segmentation (CS). Instead of running our method on the proposed non-background regions, we also test the performance of our algorithm using the full image (FI).

### 3.2 Results on iCoseg Dataset

To evaluate the performance of our method, we compare it with different image co-segmentation algorithms, e.g., MRW [Lee et al., 2015], RM [Rubio, 2012], OC [Vicente et al., 2011], CFM [Wang et al., 2013], CFO [Li et al., 2016].

Table 1 shows the accuracy of state-of-the-art image co-segmentation algorithms on iCoseg dataset. Table 2 depicts the precision and recall for the different methods. We also show some qualitative results in Fig. 2. From these figures, we have the following observations:

1. Compared to state-of-the-art image co-segmentation algorithms, our method achieves better performance in terms of segmentation average accuracy. This indicates that our graph-based coarse-to-fine image co-segmentation strategy can result in a good segmentation. MRW [Lee et al., 2015] and [Li et al., 2016] are strong competitors, and they outperform our method for some object classes, e.g., ‘elephant’, ‘skating’ and ‘stonehenge2’. One reason is that the proposed non-background regions of our method is not that accurate, and
some foreground are not covered by the non-background regions, e.g., ‘elephant’ and ‘stonehenge2’. Another reason is that HSV is a simple visual feature, and is not powerful enough to segment visually different superpixels into one part, e.g., the black skirts and white pants for ‘skating’. Compared with FI, the performance for CS is improved by 2.1%. This indicates that our proposed non-background regions can get rid of some noise, and improve our graph-based segmentation algorithm. On the other hand, the improvement of CS over IS is more significant, and it is 8.0%. IS attempts to separate an object from its background in a single image, but without consideration of multiple images, the definition of ‘object’ is ambiguous. For instance, given the single image of ‘baseball’, it is not clear whether the object should consist of the people in white only or all the people. Thus, IS may fail to delineate desired objects, especially in the cases of several objects. In contrast, CS overcomes the ambiguity, by extracting regions that occur repeatedly across images.

Comparing with MRW [Lee et al., 2015] in terms of precision and recall, our method CS achieves much better performance for precision with 7.4%, but slightly worse performance for recall with 0.7% decrease. These results show that our CS can generate relative smaller foreground than MRW, but these regions are more precise.

The qualitative results are demonstrated as the top-rows in Fig. 2. The results show that our CS can successfully co-segment images for simple background (e.g., ‘balloon’), complex background (e.g., ‘alaskan bear’), single object (e.g., ‘panda’) and multiple objects (e.g., ‘skating’). In general, better performance is achieved for single object segmentation with simple background. In cases of complex background and foreground, e.g., ‘skating’ and ‘Stonehenge2’, our segmentation performance is unsatisfactory, and the proposed foreground and background have more noise than other cases.

### 3.3 Results on MSRC Dataset

On MSRC dataset, we compare our method to state-of-the-art image co-segmentation algorithms, e.g., MRW, DC [Joulin et al., 2010], RM [Rubio, 2012], CFM [Wang et al., 2013]. Table 3 shows the accuracy of state-of-the-art image co-segmentation algorithms on MSRC dataset and Table 4 depicts the precision and recall for the different methods. We also show some qualitative results in Fig. 2. We have the following observations:

Compared to state-of-the-art image co-segmentation algorithms, our method CS achieves comparable performance in terms of segmentation average accuracy. This indicates that our CS can result in a good segmentation for cases of complex background. [Wang et al., 2013] has better accuracy than our CS, and it outperform our method for object classes of ‘car(back)’ and ‘face’ by 7.0% and 5.7%. However, they use multiple features, and our CS achieves this performance by using HSV feature only. By using sophisticated features or multiple features, the performance can be potentially improved. On the other hand, our CS obtains the best performance for ‘car(front)’, ‘bike’, ‘cat’ and ‘cow’, and it consistently outperforms MRW, [Rubio, 2012] and [Joulin et al., 2010] except for the class of ‘plane’

Compared with FI, the performance for CS is improved by 10.2%, which is greater than iCoseg dataset. The indicates that for complex background, our proposed non-background regions can better get rid of noise. On the other hand, the improvement of CS over IS is much more significant, and it is 17.0%. The reason for the improvement of CS over IS is similar to that of iCoseg dataset. CS overcomes the ambiguity of IS by extracting regions that co-occur in all the images.

Comparing with MRW [Lee et al., 2015] in terms of precision and recall, our method CS achieves much better performance for recall, but slightly worse performance for precision. These show that our CS can segment more regions to foreground than MRW, without decreasing the precision.

The qualitative results are demonstrated as the bottom-rows in Fig. 2. The qualitative results show that our CS can successfully co-segment images for simple background (e.g., ‘cow’), complex background (e.g., ‘face’), single object (e.g., ‘face’) and multiple objects (e.g., ‘bike’). In general, bet-

---

**Table 2**: Precision and Recall of state-of-the-art image co-segmentation algorithms on iCoseg dataset.

<table>
<thead>
<tr>
<th>Class</th>
<th>MRW [Lee et al., 2015]</th>
<th>MRW [Lee et al., 2015]</th>
<th>P</th>
<th>Out CS</th>
<th>R</th>
<th>Out CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Balloon</td>
<td>94.0</td>
<td>95.2</td>
<td>79.4</td>
<td>91.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Baseball</td>
<td>84.5</td>
<td>73.6</td>
<td>87.7</td>
<td>74.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Bear</td>
<td>75.9</td>
<td>98.5</td>
<td>76.8</td>
<td>97.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Elephant</td>
<td>71.9</td>
<td>97.1</td>
<td>77.2</td>
<td>74.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Ferrari</td>
<td>81.4</td>
<td>92.4</td>
<td>98.6</td>
<td>68.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Gymnastics</td>
<td>99.6</td>
<td>75.4</td>
<td>97.2</td>
<td>88.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Kite</td>
<td>95.2</td>
<td>94.6</td>
<td>89.7</td>
<td>83.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 Kite panda</td>
<td>94.5</td>
<td>93.1</td>
<td>90.8</td>
<td>98.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Liverpool</td>
<td>56.2</td>
<td>79.4</td>
<td>66.8</td>
<td>96.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 Panda</td>
<td>95.4</td>
<td>71.5</td>
<td>86.7</td>
<td>93.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 Skating</td>
<td>95.1</td>
<td>93.6</td>
<td>85.7</td>
<td>91.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 Statue</td>
<td>99.1</td>
<td>77.4</td>
<td>98.7</td>
<td>83.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 Stonehenge</td>
<td>74.0</td>
<td>97.9</td>
<td>88.7</td>
<td>92.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 Stonehenge2</td>
<td>82.3</td>
<td>96.6</td>
<td>89.1</td>
<td>86.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 Taj Mahal</td>
<td>80.8</td>
<td>92.8</td>
<td>89.2</td>
<td>95.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Average**: 81.1 | 89.1 | 87.4 | 88.2

**Table 3**: Accuracy of state-of-the-art image co-segmentation algorithms on MSRC dataset.

<table>
<thead>
<tr>
<th>P</th>
<th>R</th>
<th>Class</th>
<th>MRW [Lee et al., 2015]</th>
<th>MRW [Lee et al., 2015]</th>
<th>P</th>
<th>Out CS</th>
<th>R</th>
<th>Out CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 car(front)</td>
<td>65.9</td>
<td>87.6</td>
<td>87.3</td>
<td>84.3</td>
<td>86.0</td>
<td>66.6</td>
<td>88.7</td>
<td></td>
</tr>
<tr>
<td>2 car(back)</td>
<td>52.4</td>
<td>85.1</td>
<td>92.7</td>
<td>81.7</td>
<td>64.7</td>
<td>66.8</td>
<td>85.7</td>
<td></td>
</tr>
<tr>
<td>3 bike</td>
<td>62.4</td>
<td>63.3</td>
<td>74.8</td>
<td>70.1</td>
<td>61.6</td>
<td>60.7</td>
<td>76.2</td>
<td></td>
</tr>
<tr>
<td>4 cat</td>
<td>77.1</td>
<td>74.4</td>
<td>88.3</td>
<td>82.2</td>
<td>83.7</td>
<td>71.6</td>
<td>90.1</td>
<td></td>
</tr>
<tr>
<td>5 cow</td>
<td>80.1</td>
<td>81.6</td>
<td>89.7</td>
<td>93.1</td>
<td>92.4</td>
<td>86.2</td>
<td>93.2</td>
<td></td>
</tr>
<tr>
<td>6 face</td>
<td>76.3</td>
<td>84.3</td>
<td>89.3</td>
<td>81.0</td>
<td>74.8</td>
<td>73.2</td>
<td>83.6</td>
<td></td>
</tr>
<tr>
<td>7 plane</td>
<td>77.0</td>
<td>73.8</td>
<td>87.3</td>
<td>93.0</td>
<td>67.8</td>
<td>61.1</td>
<td>85.1</td>
<td></td>
</tr>
</tbody>
</table>

**Average**: 70.8 | 78.6 | 87.1 | 83.7 | 76.0 | 69.2 | 86.2

**Table 4**: Precision and Recall of state-of-the-art image co-segmentation algorithms on MSRC dataset.

<table>
<thead>
<tr>
<th>P</th>
<th>R</th>
<th>Class</th>
<th>MRW [Lee et al., 2015]</th>
<th>MRW [Lee et al., 2015]</th>
<th>P</th>
<th>Out CS</th>
<th>R</th>
<th>Out CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 car(front)</td>
<td>99.3</td>
<td>68.4</td>
<td>95.7</td>
<td>79.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 car(back)</td>
<td>94.4</td>
<td>68.3</td>
<td>85.8</td>
<td>85.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 cat</td>
<td>57.2</td>
<td>89.1</td>
<td>64.3</td>
<td>76.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 cow</td>
<td>74.1</td>
<td>53.5</td>
<td>83.0</td>
<td>80.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 dog</td>
<td>95.6</td>
<td>77.1</td>
<td>91.1</td>
<td>83.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 plane</td>
<td>66.1</td>
<td>58.4</td>
<td>57.2</td>
<td>72.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 sheep</td>
<td>53.1</td>
<td>88.7</td>
<td>59.5</td>
<td>78.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Average**: 77.4 | 71.9 | 56.7 | 78.5
ter performance is achieved for single object segmentation with simple background. In cases of complex background and foreground, e.g., ‘car(front)’ and ‘airplane’, the proposed foreground and background have more noise then other cases. Therefore, our segmentation performance is unsatisfactory.

4 Conclusions

In this work, we propose a novel unsupervised coarse-to-fine co-segmentation (CFC) framework, which introduces graph-based image co-segmentation based on intra and inter rank constraints, and also utilizes the coarse foreground and background proposals to construct a robust original similarity graph. Experiments on two commonly used image datasets (iCoseg and MSRC) demonstrate that our method achieves comparable or even better performance than state-of-the-art methods by using only HSV feature.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities (Grant No. ZYGX2014J063, No. ZYGX2014Z007) and the National Natural Science Foundation of China (Grant No. 61772116, No. 61502080, No. 61632007, No. 61602049).

References


