2013 IEEE International Symposium on Multimedia

Unsupervised Co-segmentation of Complex Image Set via Bi-harmonic Distance governed Multi-level Deformable Graph Clustering

Jizhou Ma[*], Shuai Li[*‡], Aimin Hao [*], and Hong Qin[†]
[*]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China*
[‡]*ls@vrlab.buaa.edu.cn*
[†]*Department of Computer Science, Stony Brook University (SUNY Stony Brook), Stony Brook 11794, USA*

***Abstract*—Despite the recent success of extensive co-segmentation studies, they still suffer from limitations in accommodating multiple-foreground, large-scale, high-variability image set, as well as their underlying capability for parallel implementation. To improve, this paper proposes a bi-harmonic distance governed flexible method for the robust coherent segmentation of the overlapping/similar contents co-existing in image group, which is independent of supervised learning and any other user-specified prior. The central idea is the novel integration of bi-harmonic distance metric design and multi-level deformable graph generation for multi-level clustering, which gives rise to a host of unique advantages: accommodating multiple-foreground images, respecting both local structures and global semantics of images, being more robust and accurate, and being convenient for parallel acceleration. Critical pipeline of our method involves intrinsic content-coherent measuring, super-pixel assisted bottom-up clustering, and multi-level deformable graph clustering based cross-image optimization. We conduct extensive experiments on the iCoseg benchmark and Oxford flower datasets, and make comprehensive evaluations to demonstrate the superiority of our method via comparison with state-of-the-art methods collected in the MSRC database.**

***Keywords*-Unsupervised Co-segmentation; Bi-harmonic Distance; High-variability Image Set; Discriminative Clustering.**

I. Introduction

The co-occurrence of common objects (or overlapping contents) in image group contains implicit supervision information. Therefore, image co-segmentation has been gaining momentum in recent years, which has widespread applications in object recognition, video segmentation, pathology identification, image set based modeling, and other web-scale applications.

Different co-segmentation methods developed under different motivations may vary in the definition and utilization of the grouping feature cues. Based on the central idea of [1], many progressively improved co-segmentation methods have been proposed by incorporating different technical insights such as Markov random field (MRF) [1]–[4], discriminative clustering [5], [6], sub-modular optimization [7], linear programming relaxation [8], subspace clustering [9], dual-decomposition [2], anisotropic diffusion [7], etc. However, in spite of their recent success, certain difficulties still prevail and need to be resolved for more intrinsic unsupervised image co-segmentation. Specifically, the common challenges existed in most of state-of-the-art methods are documented as follows.

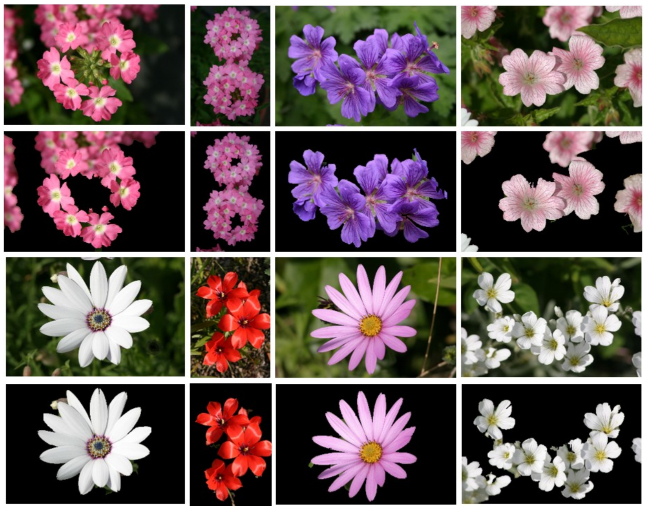


Figure 1: Our multi-class co-segmentation results for high-variability flower image set. The first and the third rows are the original image set. The second and the fourth rows are the co-segmentation results.

First, most of the state-of-the-art methods still lack enough flexibility and robustness to accommodate the complex cases due to the excessive reliance on low-level and local similarity metric such as local color consistency and texton-consistency. For example, the common objects occurring across images may vary in shape, color, scale, noise, stain, occlusion, and local deformation.

Second, existing unsupervised methods usually employ generative probabilistic model to learn labeling of co-occurring objects via iterative refinement. It gives rise to computationally intensive optimization problems, and the situation is even worse for large-scale image set. Besides, prior knowledge based learning methods usually suffer from the sophisticated tuning of the underlying classifier parameters.

Third, despite extensive co-segmentation methods, most

978-0-7695-5140-1/13 $26.00 © 2013 IEEE
DOI 10.1109/ISM.2013.16

of them rely on inter-twined computation across images, which are hard to parallelize the time-consuming tasks. From the practical point of view, a more computation-independent co-segmentation method together with its GPU acceleration is urgently needed.

To tackle the aforementioned challenges, we focus on a novel unsupervised co-segmentation method by resorting to bi-harmonic distance metric [10] governed deformable graph clustering. As illustrated in Fig. 2, we exploit and integrate the advantages of many technical elements such as superpixel based over-segmentation, anisotropic heat diffusion, and discriminative clustering. Specifically, the salient contributions can be summarized as follows:

- We define an intrinsic structure-aware bi-harmonic distance metric to measure the affinity of topology-free image contents by simultaneously integrating the pixel-level Fast Retina Keypoint (FREAK) features, superpixel-level color-consistency, and global topological information of the image contents, which facilitates the intrinsic depiction of local structure, the robust representation of intra-image global relations, and the discriminative relationship measurement of the cross-image overlapping/similar contents in a global setting.
- We propose a bottom-up co-segmentation method by integrating the intrinsic bi-harmonic distance metric into a newly designed multi-level deformable graph clustering based optimization framework, which can accommodate robust and flexible co-segmentation of multiple-foreground, large-scale, and high-variability heterogeneous image sets in an unsupervised way.
- We design image-wise parallel algorithms for most of the time-consuming tasks involved in our method, hence, the co-segmentation result can be efficiently obtained through an independent complementary optimization step to jointly merge the image-wise outputs, which collectively are very applicable to multi-thread implementation based on CPU or GPU.

II. Related Work

A. Content-coherent Measurement

Local feature descriptors. Most of earlier image co-segmentation methods [1]–[3] take pixels as basic feature elements, which usually suffer from computation intensiveness and noise sensitiveness. Considering the cross-image variability, the most intuitive histogram statistics about color, texture or frequency distribution are commonly adopted to serve as local feature descriptors. To improve, Joulin et al. [5], [6] proved that the SIFT descriptor is more robust to depict the cross-image coherency during co-segmentation. For the analogous purpose, Rubio et al. [4] exploited the advantage of Histogram of Oriented Gradients (HOG) descriptors in graph matching based foreground co-segmentation. Besides, to better accommodate the illumination variability, Glasner et al. [8] proposed region contour based descriptors by only considering the contributions of exterior boundary of local feature element. However, it is inevitable to reduce the distinguishability of the descriptor.

Affinity measurement. Some initial work [1]–[3] models the intra-image affinity measurement in terms of Markov Random Field (MRF). Rubio et al. [4] proposed a M-RF based multi-scale model to encode the graph matching results into inter-image information, which can handle more complex situations such as the high-variability of different viewpoints, illuminations, deformations, and poses. However, it is not applicable to large-scale applications. Aiming at improving the adaptivity and efficiency of large-scale co-segmentation problems, some simple but effective distance metrics such as $Euclideandistance$, $\chi^2 distance$ are also adopted to construct cross-image affinity matrix. Specifically, some recent studies resort to some intrinsic and smart distance metrics to handle more complex cases, including anisotropic heat diffusion distance [11], commute-time distance [4], [12], geodesic distance [13], which are more meaningful and informative for robust co-segmentation of objects with deformation, occlusion, noise perturbation, and flexible global relations.

B. Joint Segmentation Model

Co-clustering based methods. Co-clustering based methods can be comfortably [5], [6], [8], [9], [14] employed for image co-segmentation by modeling the problem with a set of intra-graph and a synergetic inter-graph as a global constraint to guarantee the content-coherency. To efficiently solve the co-clustering model, Glasner et al. [8] formulated the problem as a quadratic semi-assignment problem, and Joulin et al. [5] proposed a discriminative clustering framework [15] by integrating additional graph based global constraints into the generic clustering model to ensure the cross-image segmentation coherency. Meanwhile, by relaxing the discriminative clustering problem to a continuous convex optimization problem, Joulin et al. [6] can achieve the joint segmentation of dozens of images.

Learning based methods. Learning based methods generally employ manually-specified prior or unsupervised generative probabilistic models to learn labeling of co-occurring objects via an iterative refinement. For example, Vicente et al. [16] conducted co-segmentation through pair matching of pre-segmented patches by training a random forest regressor from the ground truth. Kim et al. [17] used an iterative self-learning approach to obtain a foreground representation model via the appearance statistics of mid-level elements, which facilitates the automatic label assignment for objects irregularly-occurring in an image set.

Optimization based methods. Optimization based methods solve the co-segmentation problem by maximizing/minimizing a well-designed energy function. For example, Joulin et al. [6] extended their discriminative clustering

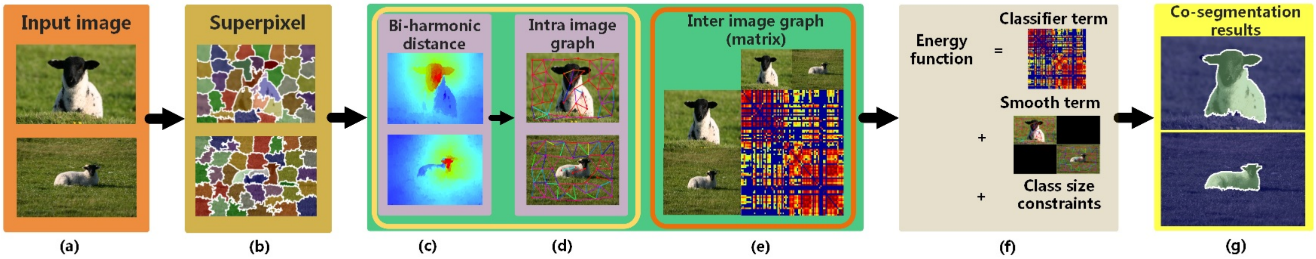


Figure 2: The flowchart of our method. (a) Input image set; (b) Super-pixels based segmentation; (c) Bi-harmonic distance based similarity measurement; (d) Bi-harmonic distance governed intra-image graphs; (e) Bi-harmonic distance governed inter-image graph, which is represented via global Laplacian matrix; (f) Energy function definition; (g) Discriminative graph clustering based co-segmentation results.

based work [5] for more favorable multi-foreground co-segmentation by defining an objective energy function comprising an intra-image coherency term, an inter-image affinity term, and a global constraint term. And they designed an Expectation Maximization (EM) algorithm to solve it. While Rubio et al. [4] constructed their energy function with a pixel-wise coherency term, a region-wise coherency term, a region-pixel relationship based scale-energy term enforce, and a cross-image region matching based energy term.

III. Method Overview

This paper focuses on an unsupervised co-segmentation method of large-scale heterogeneous image groups by extending and integrating the notions of super-pixel over-segmentation [18], manifold bi-harmonic distance definition [10], anisotropic diffusion based co-segmentation [7] into the powerful discriminative clustering framework [6]. As shown in Fig. 2, the involved steps are briefly described as follows:

Over-complete segmentation. Conduct super-pixel based over-complete segmentation for each image in the image set.

Manifold construction. Construct 3D mesh for the underlying manifold comprising super-pixels. The x-axis and y-axis coordinates encode the super-pixel's spatial information, and the z-axis coordinates embody the intensity/color properties of super-pixels.

Intra-image bi-harmonic distance definition. Define the bi-harmonic distance metric for intra-image super-pixels by employing the discrete Laplace-Beltrami operator to conduct differential analysis over the constructed manifold mesh.

Intra-image graph generation. Generate an intra-image graph for each image to represent the global relations of super-pixels. The super-pixels serve as graph nodes and their bi-harmonic distances serve as graph edge weights.

Inter-image graph generation. Generate an inter-image graph for an image set to represent the super-pixel affinity. All the super-pixels from the image set serve as graph nodes and their feature differences serve as graph edges. Meanwhile, define the bi-harmonic distance metric over the inter-image graph.

Bi-harmonic distance governed discriminative graph clustering. Define the objective energy function to encode the super-pixels' similarity in feature space (the inter-image graph), super-pixels' coherency in image space (intra-image graphs), and the uniform co-segmentation constraints. And the co-segmentation results can be obtained by minimizing the energy function with an expectation-Maximization (EM) method.

IV. Super-pixel based Manifold Mesh

Bi-harmonic distance metric [10] has achieved great success in geometry processing because of its built-in advantages such as being informative, multi-scale, robust, parameter-free, and isometric-deformation invariant. However, it remains difficult to directly define the powerful intrinsic distance metric over the images due to the following reasons: (1) 2D images comprising regular pixels have regular structure, and are both topology-free and boundary-free without any intuitive geometric meaning. (2) It is impractical to directly employ the pixel as a basic building block towards meaningful differential analysis, since the pixel-level Laplacian matrix does not support multi-scale functionality that is highly demanded in any novel shape descriptors.

We elaborate our novel intrinsic distance metric on a manifold space enabled by the construction of super-pixels. In order to guarantee the mesh regularity of the succeedingly-constructed manifold, we employ the SLIC method [18] to conduct relatively uniform segmentation for super-pixels. Meanwhile, to respect the anisotropic property reflected in original color space of the image, we inherit the anisotropy by taking the average intensity of each super-pixel as its third dimensional coordinate in 3D space. Therefore, with each super-pixel geometric center serving as a 3D vertex, the manifold mesh corresponding to each image can be constructed by means of Delaunay triangulation. Specifically, the z-axis difference between two neighboring vertices can be defined as $d_Z(v_a, v_b) = d_D(Desc(v_a), Desc(v_b))$, where $Desc$ represents a certain kind of super-pixel-level descriptor, for example, gray value, color, SIFT, texture pattern, etc. For the simplest case, an image can be transformed to an elevation map (Fig. 3) by representing z-axis with the average gray value of each super-pixel ($Desc(v) = Gray(v)$), thus $d_D(Desc_a, Desc_b) = Desc_a - Desc_b$. In this paper, we adopt $3D$ CIE Lab color properties and FREAK features [19] as descriptors. Since d_D is a distance function and its formulation depends on the adopted descriptor type, Hamming distance is favorable for FREAK descriptors, while Euclidean distance is appropriate for color-like descriptor vectors.

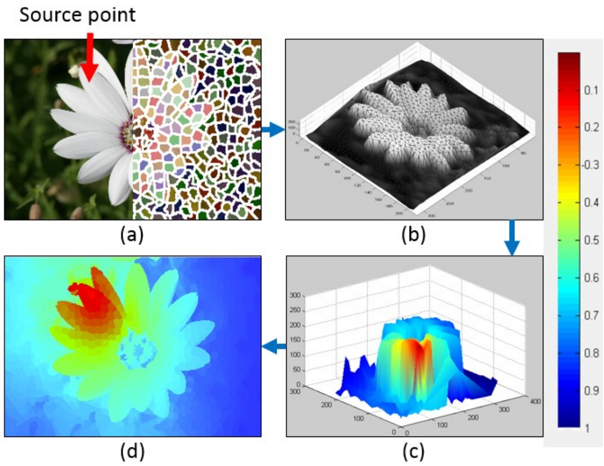


Figure 3: Illustration of bi-harmonic distance definition in an image space. (a) Super-pixels based over-complete segmentation; (b) Manifold mesh construction; (c) Bi-harmonic distance distribution over the manifold with red arrow labeled vertex as anchor; (d) Bi-harmonic distance distribution on the corresponding 2D image.

V. Bi-harmonic Distance Based Multi-level Graph

Bi-harmonic distance is built on Riemannian manifold, which can be defined using the eigenvectors and eigenvalues of the Laplace-Beltrami matrix as

$$d_B(x,y)^2 = \sum_{k=1}^{\infty} \frac{(\Phi_k(x) - \Phi_k(y))^2}{\lambda_k^2}, \tag{1}$$

where $\Phi_k(x), \lambda_k$ are respectively the eigenfunctions and eigenvalues of the positive definite Laplace-Beltrami matrix.

A. Intra-image Bi-harmonic Distance Definition

With the vertex set of the manifold mesh denoted by $P = \{p_1, p_2, ..., p_n\}$, we define the bi-harmonic distance metric via discrete Laplacian-matrix $L = A^{-1}M$ based anisotropic heat diffusion, where A is a diagonal matrix and A_{ii} is proportional to the average area of the triangles sharing vertex p_i. And M is formulated as

$$M_{ij} = \begin{cases} \sum_k m_{ij} & \text{if } i = j \\ -m_{ij} & \text{if } p_i \text{ and } p_j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where $m_{ij} = \cot\alpha_{ij} + \cot\beta_{ij}$, α_{ij} and β_{ij} are the opposite angles of two adjacent triangles sharing the edge $p_i p_j$. In intra-image case, we use $(l_i - l_j)^2 + (a_i - a_j)^2 + (b_i - b_j)^2$ to calculate the distance of the color component that is embedded in the edge length, where (L,A,B) denotes the color descriptor value of super-pixel p.

The eigen-decomposition of M is time-consuming. Accelerated bi-harmonic distance computation [10] is proposed by taking the Green's function g_d as the pseudo-inverse of $MA^{-1}M$.

$$d_B(p_i, p_j)^2 = g_d(i,i) + g_d(j,j) - 2g_d(i,j). \tag{3}$$

Since distance in some sense is a kind of dissimilarity, we can further use the Gaussian function ($\beta = 0.35$ for all types of features in all cases) to convert the bi-harmonic distance to measure similarity.

$$w(v_i, v_j) = \exp(-\frac{d_B(p_i, p_j)}{\beta}). \tag{4}$$

Therefore, we can respectively compute the affinity for each super-pixel pair. Taking the super-pixel indicated by the red arrow as an anchor vertex, Fig. 3 demonstrates the bi-harmonic distance distribution, wherein the color ranging from red to blue means that the bi-harmonic distance is going from the near to the distant.

B. Inter-image Bi-harmonic Distance Definition

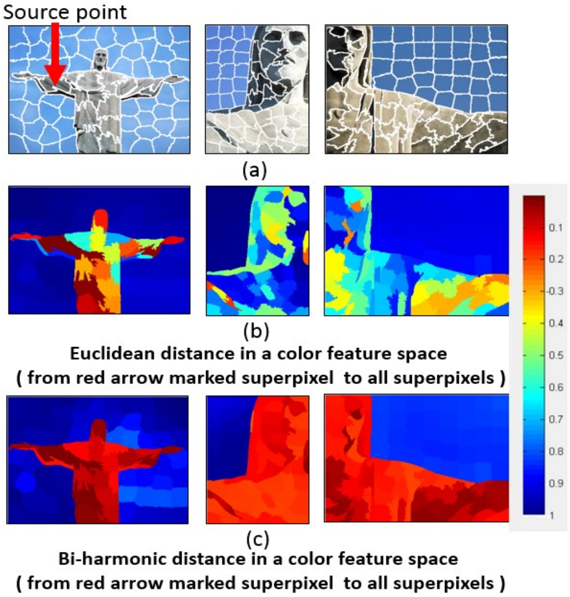
As for the bi-harmonic distance definition between inter-image super-pixels, we adopt super-pixel colors and FREAK features to redefine the Laplacian matrix M in Eq. 3. We construct a $N_P \times N_P$ dissimilarity (distance) matrix ($N_P = \sum_{i=1}^{N_I} N_P^i$, where N_I is the number of the images, N_P^i denotes the number of superpixels of the i-th image). And then we use a Gaussian function to convert the dissimilarity matrix to an affinity matrix W. Therefore, the inter-image Laplacian matrix (L_f) can be obtained as

$$L_f = E_N - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}. \tag{5}$$

L_f is a normalized Laplacian matrix [20]. E_N is the N-dimensional identity matrix. D is a diagonal matrix resulted from the row sum of W. Analogous to intra-image case, we can define the inter-image bi-harmonic similarity matrix W_B for the super-pixels in the image set.

Fig. 4 lists the comparison of bi-harmonic distance and Euclidean distance defined on three images, where the images have the same content and only vary in illumination and tone. It shows that the simple Euclidean distance metric in (L,A,B) color space is hard to provide meaningful similarity measurement for super-pixels located inside the same objects. In sharp contrast, bi-harmonic distance metric in feature space can naturally embody the indirect connectedness among super-pixels, and thus can more intrinsically reveal their similarity. Therefore, bi-harmonic distance metric can greatly facilitate the coherency measurement in a global way, which is significantly needed in image co-segmentation.

In the following sections, we respectively denote the bi-harmonic distance governed intra-image affinity matrix and the inter-image affinity matrix with W_B^i and W_B.



(b)

Euclidean distance in a color feature space

(from red arrow marked superpixel to all superpixels)

(c)

Bi-harmonic distance in a color feature space

(from red arrow marked superpixel to all superpixels)

Figure 4: Illustration of inter-image bi-harmonic distance distribution. (a) Super-pixels in original images; (b) Euclidean distance distribution with the red arrow labeled super-pixel as anchor; (c) Corresponding bi-harmonic distance distribution.

C. Multi-level Graph Construction

Based on the defined bi-harmonic distance governed affinity matrix W_B^i and W_B, we can respectively construct intra-image graphs and a fully-connected inter-image graph, where the super-pixels serve as the graph nodes and the bi-harmonic distances serve as graph edge weights. When constructing intra-image graph for each image, as shown in Fig. 2(d), we only establish edge connections for the super-pixels satisfying one-ring neighboring relations. As for the fully-connected inter-image graph, to exclude the low-probability matching relations and simplify the redundancy relations, the edge is pruned in advance if the affinity between the corresponding super-pixel pair is falling within the last 20% of the entire distance range (illustrated in Fig. 2(e)).

Directly benefiting from the local-deformation-insensitive characteristics, such graph representation can accommodate more flexible changes of co-occurring objects. Taking series of someone's photos for example, the super-pixels comprising human body should preserve close affinity in spite of complex posture changes. And the corresponding graphs should also remain stable. Therefore, our bi-harmonic distance governed multi-level graphs can support the deformable relation description of image contents in some sense.

VI. Discriminative Graph Clustering based Co-segmentation

We incorporate the bi-harmonic distance governed multi-level graph into the discriminative clustering framework [6] to perform co-segmentation. Benefitting from the superiority of bi-harmonic distance governed multi-level graph, in sharp contrast to [6], we do not need to initialize the clustering algorithm with the results of [5] as prior. Here we mainly focus on how to integrate the above-documented technical elements towards unsupervised co-segmentation, for more details about discriminative clustering based multi-class segmentation, please refer to [6].

A. Energy Function Definition

We solve the discriminative graph clustering based co-segmentation problem by minimizing a well-designed energy function. The objective energy function is composed of a classifying term, a smoothing term, and an uniform-segmentation constraining term, which can be formulated as

$$\min_{\substack{y\in\{0,1\}^{N_P\times K}\\ y1_K=1_{N_P}}} [\min_{\substack{A\in\mathbb{R}^{d\times K}\\ b\in\mathbb{R}^K}} E_U(y,A,b)] + E_B(y) - H(y), \tag{6}$$

where y is an unknown $N_P \times K$ label matrix. $y(n,k)=1$ denotes the n-th super-pixel belongs to the k-th class. E_U encodes the inter-image information by way of a kernel based method. E_B and H are discriminative clustering terms, which can respectively guarantee the co-segmentation to be smooth and uniform.

For the classifying term E_U, the χ^2 kernels are usually adopted [5], [6]. To better enhance the co-segmentation coherency, we employ the bi-harmonic distance governed multi-level graph as a graph kernel. Let us suppose $X = \{x_1, x_2, x_3 \cdots x_{N_P}\}$ is a set of d dimensional feature vectors, the central idea of kernel based method is to map X to a higher dimensional Hilbert space $\boldsymbol{F}$ to improve the linear separability. It can be formulated as

$$K(m,l) = \Phi(x_m)^T \Phi(x_l), \tag{7}$$

where Φ is a feature map and $K(m,l)$ represents the kernel function. In this paper, W_B is used to serve as kernel function. And we rewrite it as $W_B = \psi(X)^T\psi(X)$ via an incomplete Cholesky decomposition [15]. Therefore, the $N_P \times d$ feature matrix X is replaced with a $N_P \times df$ feature matrix $\psi(X)$. Here we define the degree of freedom $d_f = 5$ for the color descriptor, and $df = 1000$ for the FREAK descriptor. The classifying term E_U can be defined as

$$E_U(y,A,b) = \frac{1}{N_P}\sum_{n=1}^{N_P} l(y_n, A\psi(x_n)) + b) + \frac{\lambda}{2K}||A||_F^2, \tag{8}$$

where $\psi(x_n)$ denotes the n-th row of the matrix $\psi(X)$, while l is a soft-max function towards handling multi-foreground cases:

$$l(y_n, A, b) = -\sum_{k=1}^{K} y_{nk} log\left(\frac{exp(a_k^T\psi(x_n)+b_k)}{\sum_{m=1}^{K} exp(a_m^T\psi(x_n)+b_m)}\right), \tag{9}$$

where a_k is the k-th row of A. As for the smoothing term E_B, we first employ the bi-harmonic affinity weighed intra-image graph W_B^i to facilitate the computation of each image's normalized Laplacian matrix according to Eq. 5. Then, we collect all L^i ($i \in [1, N_I]$) to form a holistic diagonal block matrix L, which can orderly encode the intra-image coherency information at the diagonal blocks. Therefore, E_B is formulated as

$$E_B(y) = \frac{\mu}{N_P} \sum_{i=1}^{N_I} \sum_{m=1}^{N_P^i} \sum_{k=1}^{K} y_{nk} y_{mk} L_{nm}. \qquad (10)$$

As for the uniform-segmentation constraining term H, its effect is to guarantee the uniformity of co-segmentation. Specifically, it is not expected to obtain certain classes, which tend to be ignored. Since an information entropy based penalty term can well push the number of the super-pixels belonging to each class to be average, we define H as

$$H(y) = -\sum_{i=1}^{N_I} \sum_{k=1}^{K} \left(\frac{1}{N_P} \sum_{n=1}^{N_P^i} y_{nk} \right) \log \left(\frac{1}{N_P} \sum_{n=1}^{N_P^i} y_{nk} \right). \qquad (11)$$

B. EM Optimization

We adopt the EM optimization method to solve the formulated combinatorial optimization problem via non-convex relaxation. In the implementation, we initialize the EM procedure randomly. And the unknown label y is relaxed to a convex probability $y = \{y \in [0,1]^{N_P \times K} | y1_K = 1_{N_P}\}$. In an alternative way, a quasi-Newton method is used as *M-step* to find the parameters (A, b) by minimizing the energy term E_U with given y; and then a projected gradient descent method is used as *E-step* to update the label y by minimizing the energy function Eq. 6. For more details about the EM optimization solver, please refer to [6].

VII. Experimental Results and Evaluation

We have implemented our method on a PC with a Geforce GTX 660 GPU, Intel Core I7 CPU and 16G memory based on C++, and MATLAB R2013a. We demonstrate the advantages of our method via the extensive experiments on the popular MSRC, Oxford flower, and iCoseg datasets. Meanwhile, iCoseg and Oxford flower datasets provide a binary ground truth about the foreground and the background, while MSRC dataset is a multi-class segmentation database.

During evaluation, we take the famous state-of-the-art unsupervised multi-foregrounds (or multi-class) co-segmentation methods (MC [6] and DS [7]) as competitors, all of which provide open-source codes. To make a complete evaluation within the limited space, we design the experiment groups by combining the subset of different capabilities in handling foreground complexity, image set scale, and image noise degree. Meanwhile, we adopt the standard accuracy measurement indicator $score = \frac{GT_i \bigcap R_i}{GT_i \bigcup R_i}$ to conduct quantitative evaluation of the co-segmentation quality. Moreover, to guarantee the rigorousness, all of the quantitative experimental data is obtained by an average of 20 times experiments.

Table I: The performance comparison of different co-segmentation methods. N: the scale of image set; K: co-segmented classes.

Class	N	K	Score (%)			Time (s)		
			Ours	MC	DS	Ours	MC	DS
White flowers	62	2	76.1	**78.8**	62.4	**109.2**	2241.2	149.1
Yellow flowers	49	2	83.9	**86.1**	82.2	**101.6**	1636.4	120.9
Flowers	10	2	**81.0**	66.1	64.1	42.9	449.8	**22.2**
Goose	25	2	**84.4**	80.4	67.3	46.6	108.7	**45.8**
Helicopter	12	2	**70.2**	34.0	10.3	29.2	567.9	**17.2**
Christ	13	2	**79.7**	78.6	56.6	**26.1**	532.4	26.2
Kendo	30	2	**85.0**	82.5	54.0	**48.8**	1328.8	50.0
Sheep	12	2	73.4	**73.8**	63.9	22.1	426.5	**20.7**
Bench	9	2	**66.8**	57.2	52.7	**11.1**	288.0	17.3
House	5	2	**70.7**	60.9	19.5	13.8	137.0	**8.3**
Brand	5	2	**86.4**	79.0	57.1	13.4	216.0	**9.0**
Cattle	5	3	**71.5**	62.5	66.1	12.2	224.0	**7.0**
Plane	8	3	**50.0**	41.9	39.8	36.4	257.7	**15.2**
Ferrari	9	4	**68.6**	66.2	65.0	25.8	432.1	**16.8**
Base ball	25	4	**63.8**	58.0	49.4	75.2	3216.2	**52.8**
Monk	17	5	**73.0**	57.8	72.4	105.5	3384.0	**34.2**

A. Parameter Setting of Different Methods

Considering fair comparison and the tolerable runtime memory cost of different methods, for the same scale image set, we resize all the images to be 256^2 pixels in our method and DS method [7] while the size of 128^2 pixels is utilized in MC method [6]. The number of super-pixels is set to be 80 in all the experiments. Empirically, taking "80" as the basic number of super-pixels can ignore the details and such number is not too big for the resized images, while the other numbers of super-pixels with the same order of magnitude may work as well. And the weight μ in the smoothing term E_B is always set to be 10^4 in our method. Gaussian function parameter β equals to 0.35. Since MC involves two types of features: SIFT and color, we select the better results obtained from the two types of features for comparison. Besides, MC needs to be initialized with [5]'s results. However, when the number of image set is larger than 40 or the image set contains heavy noisy, MC initialization may not work due to the intolerable time cost of [5]. In such cases, MC can only be initialized with original image set. As for DS method, since its central idea is to orderly select the color-coherency area as large as possible via a greedy algorithm, DS is color-sensitive. Therefore, suppose we set the co-segmented object classes to be K, we choose DS's best results corresponding to the parameter range $[K, 2K]$ for comparison. Moreover, to further compensate for DS's color sensitivity, we allow adjusting its Gaussian parameter within the range $0.25 \leq \beta \leq 0.6(default)$ when DS gives completely wrong results.

B. Single Foreground Co-segmentation

In Fig. 5((a),(b),(c)), three groups of single foreground co-segmentation results ($K = 2$) are respectively obtained using our method (the second row in each group), MC method (the third row in each group), and DS method (the fourth row in each group). According to the quantitative accuracy and timing cost measurement in Table I, our method outperforms other methods in accuracy. And DS is sometimes more efficient than ours, however, DS has poorest accuracy because the similar objects in different images always tend to vary (at least slightly) in color, illumination, and tone. In contrast, our method and MC can alleviate this problem because the energy term gives rise to soft segmentations. Besides, the color coherency of super-pixels also facilitates further enhancement of the results due to the omission of some trivial details. Meanwhile, since MC needs to compute chi-square kernel and it optimizes the energy function based on dense feature points (a small-scale image may contain thousands of feature points), the timing cost and memory consumption increase exponentially with the growth of the image or image set scale. Therefore, MC is unsuitable for large-scale image sets in practice.

When comparing with DS and MC, our method has clear advantages. our method not only inherits all the advantages from discriminative clustering framework but also improves its efficiency via the super-pixel based bottom-up clustering design. About 90% timing cost of our method comes from the final optimization step, while the rest 10% is from Matlab and CUDA based parallel computation. Since the convergence speed of EM step partially depends on initialization and the initialization is assigned randomly, the timing cost of our method has no obvious linear relationship with the image set scale. For example, the image number of ”Goose” group is twice the number of ”Flowers” group, but their efficiency test is rather close.

C. Multiple Foreground Co-segmentation

Although both MC and DS claim to support multi-foreground co-segmentation. Benefitting from the global structure awareness of the bi-harmonic distance based similarity metric, as the results shown in Fig. 1 and Fig. 5((d),(e)), our method can better facilitate multiple foreground co-segmentation. For multiple types of different flowers shown in Fig. 1, our bi-harmonic distance based similarity metric can effectively distinguish the flowers from the green background. In particular, although both our method and MC method are based on a similar framework, MC performs badly for ”Helicopter” group, which has very similar color distribution in each image. Our method has a better result because the bi-harmonic distance respectively enhances the relationship between foreground and different parts belonging to the background. Since the testing image datasets lack corresponding rigorous multi-foreground ground truth (e.g., MSRC database takes a brown cattle and a black cattle as a same class). we increase the number of class K and quantitatively evaluate their ability in multi-foreground co-segmentation by respectively comparing the segmented foreground with their corresponding ground truth. As shown in Table I, our method is the most accurate one for all the $K > 2$ cases.

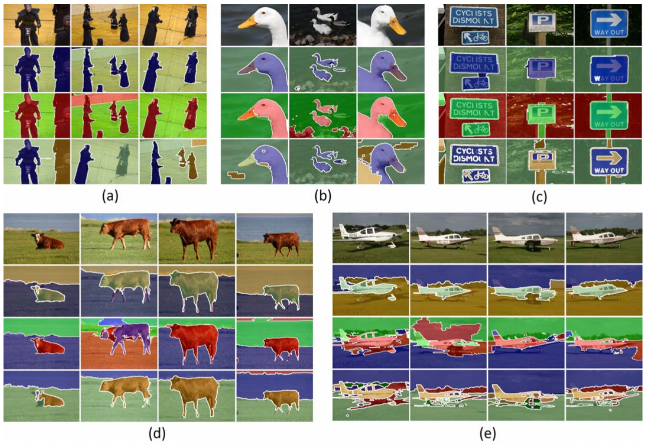


Figure 5: Co-segmentation results from different methods and their comparison over 3 single-foreground ((a), (b), (c)) and 2 multi-foreground ((d), (e)) image groups. First rows: original images; Second rows: ours; Third rows: MC [6]; Fourth rows: DS [7].

D. Noisy Image Set Co-segmentation

To verify the robustness of our method, we employ noise-perturbed ”Sheep” image group to conduct evaluation, since this image group has appropriate complexity and all the methods have close performance on its noise-free version. We randomly choose three images from the image set to respectively perturb them using 0% to 30% salt & pepper noise with 5% interval. As shown in Fig. 6(c), DS method has the worst performance. MC method can obtain the rough area of the foreground, which indicates close accuracy to ours. However, MC results are too discontinuous with many small holes that destroy the global integrity of the entire foreground, while our results can better preserve the holistic coherency thanks to the intrinsic properties of bi-harmonic distance metric. Meanwhile, Fig. 6(b) illustrates the corresponding accuracy score statistics, which indicates that our method is more robust than others. For more comparison results and more details of the pipeline, please check out the authors’ websites. http://www.vrlab.buaa.edu.cn/JizhouMa

VIII. Conclusion

In this paper, we have proposed a novel and versatile method to address a suite of research challenges in unsupervised co-segmentation for multi-foreground, large-scale, high-variability image sets. The extensive experiments and accompanying rigorous evaluation verify the superiority of our method. Specifically, the critical and novel technical elements include bi-harmonic distance metric definition of underlying manifold embedded in image, affinity measurement integrating local feature and global coherency, and multi-level deformable graph clustering, all of which also

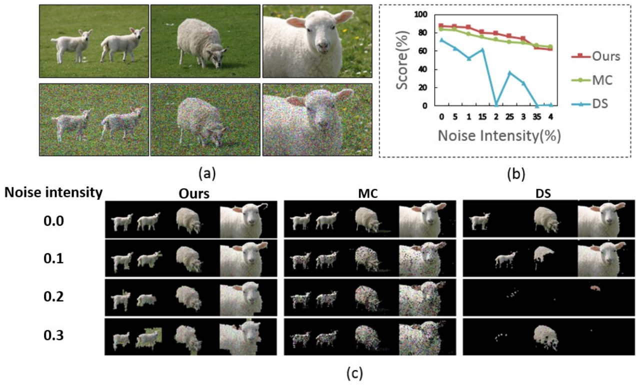


Figure 6: Co-segmentation results from different methods and their comparison over the noise-perturbed image group. (a) Three salt & pepper noise-perturbed images (randomly chosen from the 12 "Sheep" images). (b) Corresponding co-segmentation accuracy analysis. (c) Co-segmentation results.

contribute to physics-based vision, image representation, and pattern recognition collectively. Our on-going effort is geared towards finding a self-adaptive way to more intuitively determine the involved parameters. Extending our key idea to design a mid-level feature descriptor for non-rigid patch-level registration, image annotation, and image retrieval is another research direction we are exploring.

IX. Acknowledgement

This research is supported by National Natural Science Foundation of China (61190120, 61190121, 61190125, and 61300067) and National Science Foundation of USA (IIS-0949467, IIS-1047715, and IIS-1049448).

References

[1] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Co-segmentation of image pairs by histogram matching - incorporating a global constraint into mrfs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 993–1000.

[2] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *IEEE International Conference on Computer Vision*, 2009, pp. 269–276.

[3] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2028–2035.

[4] J. C. Rubio, J. Serrat, A. M. Lpez, and N. Paragios, "Unsupervised co-segmentation through region matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 749–756.

[5] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1943–1950.

[6] ——, "Multi-class cosegmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 542–549.

[7] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *IEEE International Conference on Computer Vision*, 2011, pp. 169–176.

[8] D. Glasner, S. N. P. Vitaladevuni, and R. Basri, "Contour-based joint clustering of multiple segmentations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2385–2392.

[9] R. Hu, L. Fan, and L. Liu, "Co-segmentation of 3d shapes via subspace clustering," *Eurographics Symposium on Geometry Processing*, pp. 1703–1713, 2012.

[10] Y. Lipman, R. Rustamov, and T. Funkhouser, "Biharmonic distance," *ACM Transactions on Graphics*, vol. 29, no. 3, Jun. 2010.

[11] F. Moreno-Noguer, "Deformation and illumination invariant feature point descriptor," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1593–1600.

[12] R. Behmo, N. Paragios, and V. Prinet, "Graph commute times for image representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[13] A. Criminisi, T. Sharp, and A. Blake, "Geos: Geodesic image segmentation," in *European Conference on Computer Vision*, 2008, pp. 99–112.

[14] O. Sidi, O. van Kaick, Y. Kleiman, H. Zhang, and D. Cohen-Or, "Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering," in *ACM Special Interest Group for Computer Graphics Asia Conference*, vol. 30, no. 6, 2011, pp. 126:1–126:10.

[15] F. Bach and Z. Harchaoui, "Diffrac: a discriminative and flexible framework for clustering," in *International Conference on Neural Information Processing Systems*, 2007, pp. 49–56.

[16] S. Vicente, C. Rother, and V. Kolmogorov, "Object co-segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2217–2224.

[17] G. Kim and E. P. Xing, "On multiple foreground co-segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 837–844.

[18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 2274–2282, 2012.

[19] R. Ortiz, "Freak: Fast retina keypoint," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 510–517.

[20] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 2000.