Object Co-Segmentation Based on Shortest Path Algorithm and Saliency Model

Fanman Meng, Hongliang Li, *Senior Member, IEEE*, Guanghui Liu, and King Ngi Ngan, *Fellow, IEEE*

Abstract—Segmenting common objects that have variations in color, texture and shape is a challenging problem. In this paper, we propose a new model that efficiently segments common objects from multiple images. We first segment each original image into a number of local regions. Then, we construct a digraph based on local region similarities and saliency maps. Finally, we formulate the co-segmentation problem as the shortest path problem, and we use the dynamic programming method to solve the problem. The experimental results demonstrate that the proposed model can efficiently segment the common objects from a group of images with generally lower error rate than many existing and conventional co-segmentation methods.

Index Terms—Co-saliency, co-segmentation, shortest path algorithm.

I. INTRODUCTION

IN computer vision, object segmentation plays a fundamental and important role in solving many high-level vision problems, such as object detection, object recognition and scene understanding. In recent years, a variety of image segmentation methods [1]–[5] have been proposed to deal with object segmentation problem. Due to the variety and complexity of the objects, most existing object segmentation methods focus on generating object prior through human interaction, i.e., object prior is modeled from training images [6], [7] or the users are asked to provide segmentation cues manually [8]–[10]. These methods produce huge workload for users when the number of the target images is high.

Co-segmentation aims to segment common objects from a collection of images given by the user. Compared with traditional segmentation methods, co-segmentation can accurately segment objects from images by several related images, and requires less user workload. In this paper, we focus on segmenting common objects from a set of images.

In general, a co-segmentation approach addresses the co-segmentation problem from two aspects, i.e., single image segmentation and common objects segmentation. The single image segmentation technique extracts some uniform and homogeneous regions with respect to texture or color properties, and the common objects segmentation is concerned with the segmentation of objects with similar features.

In the existing co-segmentation models, co-segmentation is commonly modeled as an optimization problem, which takes foregrounds similarity into account [11]–[21]. A Markov Random Fields (MRF) based co-segmentation method was first proposed by Rother *et al.* [11], which segmented common objects through adding the constraint of foreground similarity into traditional MRF based segmentation methods. In [11], Rother *et al.* employed L1-norm to measure foreground similarity and used trust region graph cuts (TRGC) method for energy function optimization. Since the additional constraint increases the complexity of the energy function optimization, several modifications to the constraint had been used. In the work of Mukherjee *et al.* [12], L1-norm was replaced by L2-norm. Pseudo-Boolean optimization method was used for the energy function optimization. Instead of penalizing foreground difference, Hochbaum and Singh [13] rewarded foreground similarity, which simplified the energy function optimization. In [14], Vicente *et al.* extended the foreground similarity measurement by modifying Boykov-Jolly model. Dual decomposition was employed for the energy function optimization.

Apart from MRF based segmentation method, Joulin *et al.* [16] combined discriminative clustering and spectral clustering method to perform co-segmentation. In this way, a classifier trained by spectral clustering technique and positive definite kernels was used for common objects segmentation. An interactive co-segmentation method was proposed by Batra *et al.* in [17], which segmented common objects through human interaction guided by an automatic recommendation system. Mukherjee *et al.* [18] proposed a scale invariant co-segmentation method which segmented common objects with the requirement that the rank of the matrix corresponding to foreground regions should be equal to one. Chang *et al.* [19] proposed a novel global energy term which considered both foreground similarity and background consistency. Since the energy function was submodular, the energy function can be solved effectively by the graph-cut algorithm. Vicente *et al.* [20] recently proposed a new co-segmentation method, namely

object co-segmentation, which emphasized interesting objects co-segmentation. In order to segment common objects, Vicente *et al.* selected useful features from a total of 33 features through random forest regressor. In [21], Kim *et al.* proposed a diffusion-based optimization framework which used anisotropic heat diffusion method to locate seeded points of the common objects and employed random walks segmentation method for common objects segmentation.

The existing co-segmentation methods have achieved impressive results under certain situations. However, co-segmentation still faces several challenges. The first is how to segment common objects under similar background. For example, the background of the images may be all *meadows* which are similar with each other on many features. The second challenge is how to segment common objects when the common objects share similar contours or shapes. The third is how to segment common objects from a large number of original images, which makes the problem more expensive to compute.

In this paper, we propose a new co-segmentation model which consists of three steps. The first is to segment the original images into a number of local semantic regions, which is achieved through combining superpixel based segmentation method, object detection method and saliency based segmentation method. In the second step, we design a digraph to represent the local region similarities according to the feature distance and the saliency map. In order to improve the saliency map, we employ co-saliency strategy to obtain more accurate saliency map. Based on the constructed digraph, the co-segmentation can be achieved by selecting a set of nodes with maximum sum of weights. Thus, we formulate the co-segmentation problem as a shortest path problem in the final step, and we use dynamic programming method to solve the problem. We evaluate our method on many groups of images. The experimental results demonstrate the effectiveness of our method.

This paper is organized as follows. Our proposed method is introduced in Section II. Experimental results are provided in Section III to support the efficiency of our proposed algorithm. Finally, in Section IV, conclusions are drawn.

II. PROPOSED CO-SEGMENTATION METHOD

The flowchart of the three steps of the proposed method is shown in Fig. 1. In what follows, we describe in detail the multiple local region generation method, the approach of graph construction, and the final shortest path searching algorithm.

A. Multiple Local Region Generation

The proposed multiple local region generation method segments the original image into a number of local regions R which consists of three subsets $R = \{R_1, R_2, R_3\}$. R_1 contains superpixels segmented by the over segmentation method. R_2 is comprised of the segmentation results obtained by saliency detection based method. For R_3, we first detect objects from the original images through object detection method. Then, the detected objects are segmented from the original images and treated as the elements of R_3.

We use the methods in [22] and [23] to obtain R_1 and R_2. In [22], the oriented watershed transform (OWT) is used to form the initial regions from image edges. Based on the initial
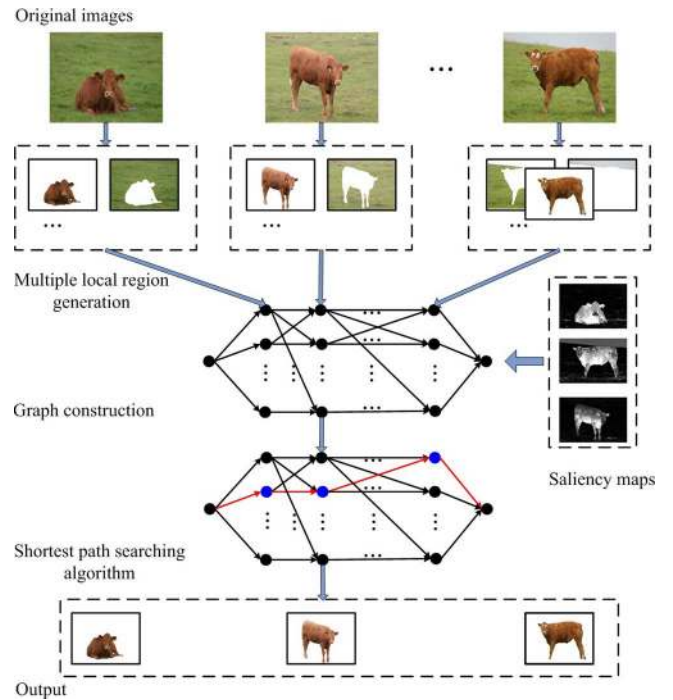

Fig. 1. Framework of the proposed method.

regions, the hierarchy of regions is constructed by the greedy graph-based region merging algorithm. The hierarchy of the regions is then treated as an ultrametric contour map (UCM). By setting different thresholds (the scale k) on the UCM, we obtain a series of segmentations to form R_1. In this paper, we set $k = 200, 150$ and 100. The saliency based object segmentation method in [23] uses global contrast based saliency detection method for object detection. The detected object is then segmented using Grabcut method. The segmented regions comprise R_2.

For R_3, we employ object detection method in [24] for detection. The method in [24] presents an objectness measure for a window to detect objects from images. Four cues are combined to measure the window such as multi-scale saliency, color contrast, edge density and superpixels straddling. By learning the generic objects in a Bayesian framework, the object in new images can be located. In the proposed method, we select two cues for evaluating the windows, i.e., color contrast and superpixel straddling. Furthermore, we improve the method of calculating color contrast in [24] by using superpixel instead of pixel. The detected objects are segmented through combing the superpixels inside the windows. Then, the obtained local regions comprise R_3. We also use hierarchical approach of [22] for obtaining superpixels. Here, we set $k = 200, 150, 100$ and 50. For each hierarchy, we select the objects corresponding to the windows with large scores (at most ten windows are selected) as local regions. By the proposed multiple local region generation method, the average number of regions for an image is about 29 over the images used in this paper.

Fig. 2 shows the segmentation results of the proposed multiple local region generation method. The classes of *ferrari*, *gymnastics*, and *hot balloons* are displayed. Eight local regions for each class are shown. The local regions are covered by
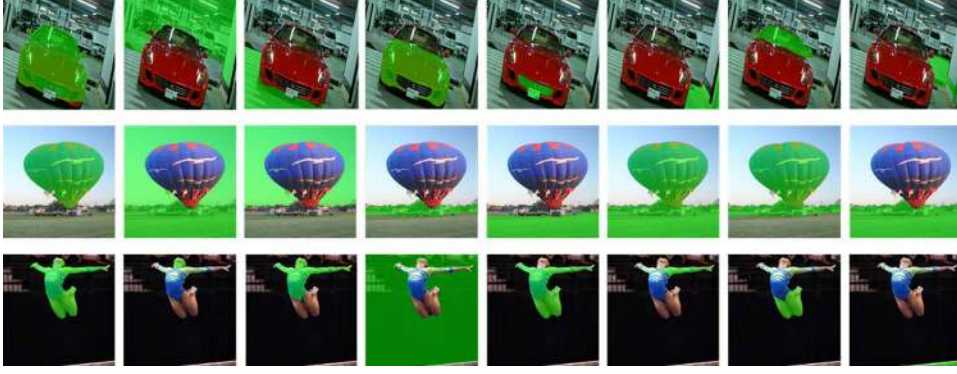
Fig. 2. Segmentation results of the proposed multiple local region generation method.

green color. We can see from Fig. 2 that the original images are divided into kinds of local regions. For example, in *hot balloons*, the "sky", "hot balloon", and "ground" are segmented as local regions. Furthermore, the unit region of "hot balloon" and "ground" is segmented as local regions. We can also see that the common objects are segmented as a local region, such as the "car" in *ferrari*.

B. Graph Construction

Based on the assumption that the common object is segmented as a local region by the multiple local region generation method, the co-segmentation can be achieved by selecting the common objects from the local regions according to their similarities. Here, we represent the similarities between local region pairs by labeled and directed graph. Assuming L_V and L_E denote the set of nodes and edge labels, a labeled and directed graph G is a 4-tuple $G = (V, E, \nu, \xi)$, where
- V is a finite set of nodes;
- $E \subseteq V \times V$ is the set of edges;
- ν is a function assigning labels to the nodes; and
- ξ the node labeling function.

Here, each node has a label v and a property f. A directed edge e corresponding to nodes v_1 and v_2 is depicted as $e = (v_1, v_2)$ (directed from v_1 to v_2). Furthermore, a weight w is assigned for each edge.

We introduce the proposed graph construction by three steps, i.e., nodes generation, edge generation and weights assignments. Assume $I = \{I_i, \ldots, I_m\}$ denotes the original image set, where m is the number of the original images. The jth local regions of I_i obtained by the multiple local region generation method is depicted as p_{ij}, $j = 1, \ldots, n_i$. n_i is the number of local regions of the image I_i. In the node generation, we first generate node v_{ij} for each local region p_{ij}, and obtain a node set $V = \{v_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i\}$. Then, we divide V into $V = \{V_1, V_2, \ldots, V_m\}$ according to image I_i, i.e., $V_i = \{v_{ij} | j = 1, \ldots, n_i\}$. Furthermore, we add tow subsets $V_0 = \{v_{01}\}$ and $V_{m+1} = \{v_{(m+1)1}\}$ into the node set for the common objects segmentation. Hence, $V = \{V_0, V_1, \ldots, V_{m+1}\}$.

After node generation, we then link any node pair v_{ij} and v_{kl} with $k = i + 1$ by the edge $e = (v_{ij}, v_{kl})$. Then, each edge $e = (v_{ij}, v_{kl})$ is assigned a weight of $w_{ij,kl}$ to represent the similarity between the local regions. A large weight will be
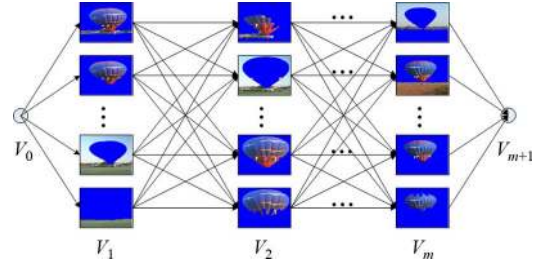

Fig. 3. Example of generated graph.

given to the similar local region pairs. The calculation of $w_{ij,kl}$ will be detailed introduced in Section II-C.

We can see that the constructed graph has the following properties which can simplify the common objects extraction:
- Node set can be classified into $m + 2$ subsets, $V_0, V_1, \ldots, V_{m+1}$, and $V_0 \cup V_1 \cup \cdots \cup V_{m+1} = V$, $V_k \cap V_l = \emptyset$, $(k \neq l)$, $0 \leq k, l \leq m + 1$.
- Each subset is called a layer. The first layer and the last layer are V_0 and V_{m+1}, respectively.
- For $\forall (u, v) \in E$, if $u \in V_k$, then $v \in V_{k+1}$.

An example of the generated digraph is shown in Fig. 3.

C. Weights Calculation

For each edge $e = (v_{ij}, v_{kl})$, we calculate the weight $w_{ij,kl}$ by

$$w_{ij,kl} = w^1_{ij,kl} + \alpha \cdot w^2_{ij,kl} \qquad (1)$$

where $w^1_{ij,kl}$ (i.e., region term) represents the region similarity between nodes v_{ij} and v_{kl}, $w^2_{ij,kl}$ (i.e., saliency term) denotes the saliency values of the two nodes, and α is scaling parameter.

1) Region Term, $w^1_{ij,kl}$: The region term represents the feature similarity between two local regions and is given by

$$w^1_{ij,kl} = d(f_{ij}, f_{kl}) \qquad (2)$$

where f_{ij} and f_{kl} are the features of the local regions p_{ij} and p_{kl}, such as color histogram or shape descriptor, and $d(\cdot, \cdot)$ denotes the distance between the two features. For the edges corresponding to the nodes in V_0 and V_{m+1}, we set $w^1_{0j,1l} = 1$ and $w^1_{mj,m+1l} = 1$.

For d in (2), the distances vary with features. We consider two features in this paper, i.e., color feature and shape feature. When color feature is considered, f_{ij} is normalized color histogram. Then, $d(\cdot,\cdot)$ is calculated by

$$d(f_{ij}, f_{kl}) = \sum_{a=1}^{\beta} \min(f_{ij}(a), f_{kl}(a)) \tag{3}$$

where β is the length of f_{ij}. For the shape representation, we employ the method [25] to obtain shape descriptor f_{ij}. The Hausdorff distance is used to calculate feature distance. Note that value 0 means the best match by Hausdorff distance, which is opposite to our assumption in the Section II-B. We use the negative value of the distance as the similarity measurement.

2) Saliency Term, $w_{ij,kl}^2$: Since the original images have similar background, we introduce the saliency term to distinguish the common objects from the similar backgrounds. The saliency term $w_{ij,kl}^2$ is calculated via the following three steps:

1) Calculate saliency map S_i for each image I_i by saliency detection method.
2) Calculate saliency value s_{ij} for each local region p_{ij}, which can be depicted by

$$s_{ij} = \left(\frac{\sum\limits_{(k,l) \in p_{ij}} (S_i(k,l))}{m_{ij}} \right) \cdot \left(\frac{m'_{ij}}{M'_i} \right) \tag{4}$$

where m_{ij} is the number of the pixels in the local region p_{ij}, m'_{ij} is the number of the salient pixels in the local region p_{ij}, and M'_i is the number of salient pixels in the I_i. The salient pixel is the pixel with salient value larger than a threshold. We set the threshold as 0.5.
3) For the edge $e = (v_{ij}, v_{kl})$, we calculate $w_{ij,kl}^2$ by

$$w_{ij,kl}^2 = s_{ij} + s_{kl}. \tag{5}$$

Note that in step 2), we use the mean saliency values and the ratio of m'_{ij}/M'_i to obtain s_{ij}. The first term is to select local regions with large mean saliency value. Because a small local region may have large mean saliency value, which results in incomplete segmentation, we use the second term to avoid the large s_{ij} value in a small local region.

Inspired by our previous work [26], we incorporate co-saliency model to improve the saliency map S_i when color feature is considered. The co-saliency model extracts saliency maps from a group of images, where the common objects in the original images are considered as saliency objects, and the similar local regions shared by the images are considered as saliency regions. Compared with single image based saliency detection method, co-saliency model can extract more accurate saliency map because the saliency regions can be more accurately located through matching with the other images.

Assume I_i is RGB image and S_i is the saliency map obtained by single image based saliency detection method. We first quantize each color channel (R, G, and B) to have 12 values and finally obtain colors φ_{ik}, $k = 1, \ldots, b_i$, where b_i is the number
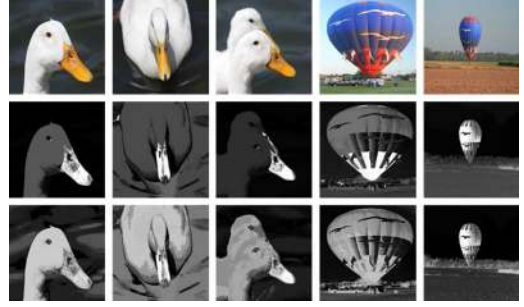

Fig. 4. Saliency maps by using [23] and our co-saliency model. First row: the original images. Second row: saliency maps by using [23]. Last row: saliency maps by using the proposed co-saliency model.

of colors. Based on S_i, we then calculate co-saliency value τ_{ik} for color φ_{ik} by

$$\tau_{ik} = r_{ik} + \eta_{ik} \tag{6}$$

where

$$\eta_{ik} = N \left(\sum_{j=1, j \neq i}^{m} \sum_{l=1}^{b_j} d(\varphi_{ik}, \varphi_{jl}) r_{jl} \right) \tag{7}$$

and r_{ik} is the mean saliency value over the pixels with the color φ_{ik}. $N(\cdot)$ is normalized function to make the values within the range of $[0,1]$. d is the distance between color vectors. Based on τ_{ik}, for any location (j,l) on the image I_k, $S_i(j,l)$ for co-saliency model can be obtained by

$$S_i(j,l) = \tau_{ik}, \text{ if } I_i(j,l) = \varphi_{ik}. \tag{8}$$

In this paper, we obtain S_i through the method in [23]. Note that in [23], the saliency maps can be obtained by two approaches, namely histogram based contrast (HC) and region based contrast (RC). Here, we select HC for saliency map calculation. The distance in (7) is given by

$$d(\varphi_{ik}, \varphi_{jl}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-|\varphi_{ik} - \varphi_{jl}|}{2\sigma^2} \right) \tag{9}$$

where φ_{ik} and φ_{jl} are RGB color vectors and $|\cdot|$ is L1-norm. In our experiments, $\sigma = 5$. In order to reduce noisy saliency in (7), we employ the smoothing procedure proposed in [23] to refine the saliency value.

From (6), we can see that our co-saliency model consists of two terms. One is r_{ik}, which is the saliency value of single image based saliency map. The other is η_{ik}, which is calculated according to the saliency maps of the rest images. Note that we take saliency maps of the rest images into account for calculating η_{ik}. The benefit of employing saliency maps of the rest images is to avoid the influence of the background similarity. For example, the saliency value of "meadow" is commonly small in single image based saliency map. Thus, the value of r_{ik} is small, which leads to a small η_{ik} according to (7).

The results by the proposed co-saliency model are displayed in Fig. 4, where the first row shows the original images. The saliency maps by the method in [23] are displayed in the second
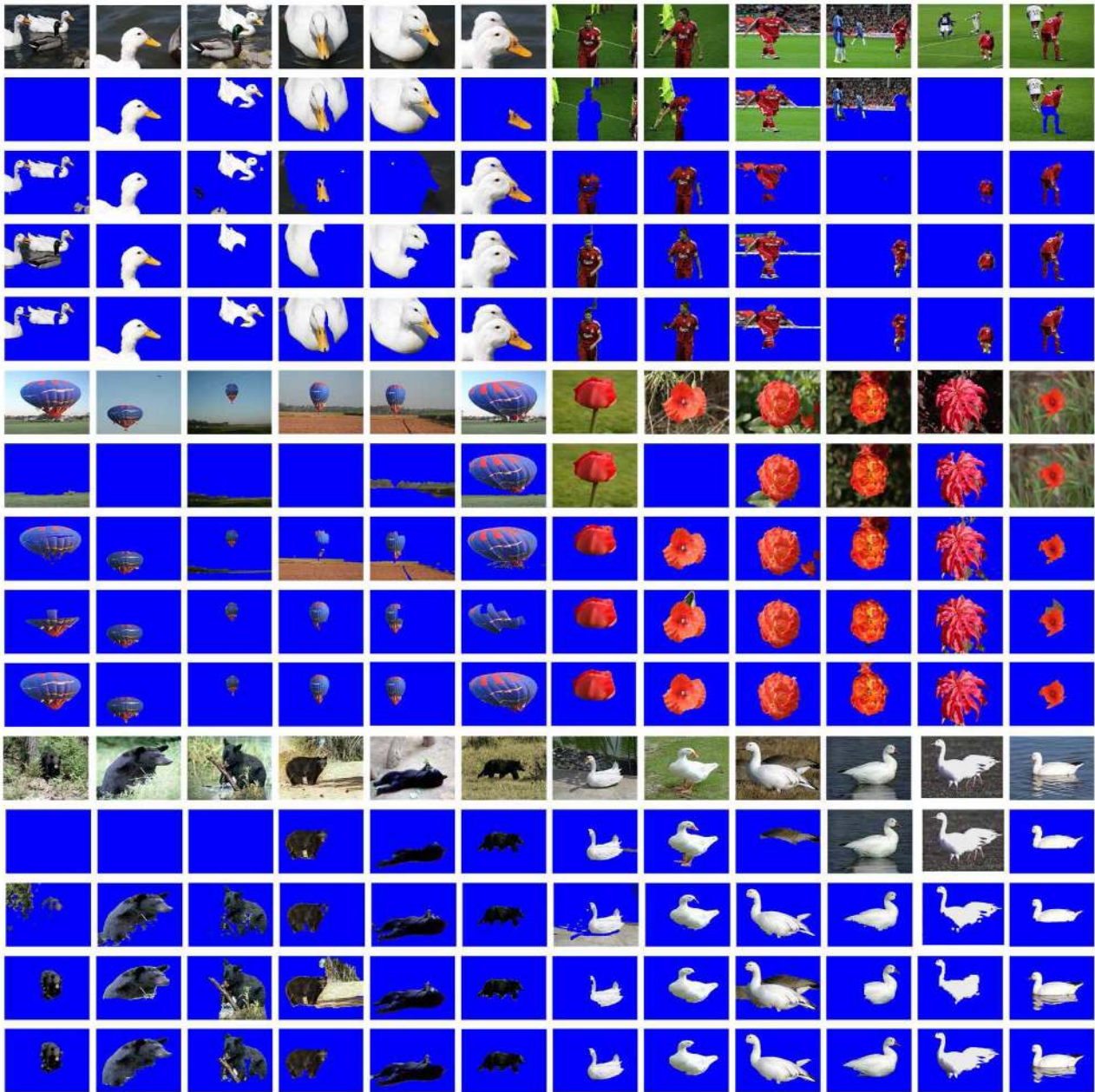
Fig. 5. Segmentation results of [16], [20], [21], and the proposed method by considering color. These classes are *goose2*, *FC player*, *hot balloons*, *flowers2*, *bear*, and *goose1*. The rows 1, 6, 11: original images. The rows 2, 7, 12: the results for the method in [16]. The rows 3, 8, 13: the results for the method in [21]. The rows 4, 9, 14: the results for the method in [20]. The rows 5, 10, 15: the results by the proposed method.

row for comparison. The third row shows the results by the proposed co-saliency method. From Fig. 4 we can see that the proposed co-saliency method improves the saliency detection. For example, in the first image the saliency value of "goose" is small for the saliency map of [23], while the saliency value in the co-saliency map is large.

D. Common Objects Segmentation

Since the common objects have both similar features (large $w^1_{ij,kl}$) and large saliency values (large $w^2_{ij,kl}$), the common objects can be segmented through searching the path that has the largest sum of the weights, which corresponds to the shortest

path searching problem. Hence, we formulate the co-segmentation as identifying the shortest path in the generated digraph structure. Based on the special properties of our constructed digraph in Section II-B, we can treat each layer as a state with time t. Thus, the shortest path problem is a dynamic decision problem, and the shortest path can be efficiently obtained by dynamic programming method.

III. EXPERIMENTAL RESULTS

In this section, we verify the proposed co-segmentation algorithm on many groups of images. The subjective and objective assessments of the segmentation results are reported.

A. Co-Segmentation Results

1) Test Images Dataset: In order to completely verify our method, we collect image groups commonly used in co-segmentation, such as ICoseg database given by [17]. In addition, we select image classes from several well-known image databases, such as MSRC database, Caltech database, and ETHZ shape database. Note that some classes have similar backgrounds, such as *cow* in MRSC, and some classes have common objects that share similar contours, such as *applelogos* in ETHZ shape database.

2) Evaluation Using Color Features: We first show the co-segmentation results of the proposed method using color feature. The results are shown in Fig. 5, where six classes are displayed. For each class block, there are fifth rows. The first row shows the original images. The results of the proposed method are shown in the last row. We can see that the proposed method successfully segments common objects from these images. For example, the "FC players" are segmented from original images which have similar backgrounds.

We also compare our method with the exiting co-segmentation methods in [16], [21], and [20]. Joulin *et al.* in [16] proposed co-segmentation model by using discriminative clustering and spectral clustering method. In [16], a supervised classifier trained from a label of the images corresponds to a separation. The label leading to the maximal separation of the two classes is the co-segmentation result. The corresponding optimization problem can be solved by relaxing to a continuous convex optimization problem. In our experiment, we use the source code given by the authors.[1] Chi-square kernel is selected as the kernel. The method of [22] is used for superpixels segmentation (by setting $k = 100$). The SIFT feature is used to represent local regions for the classes from MSRC dataset. For the other classes, the color histogram is employed as the feature. The segmentation results of the method in [16] are shown in the second row of each class block in Fig. 5. It is seen that the common objects are accurately segmented from the classes, such as *goose2* and *goose1*. As the training of the classifier can be affected by the similar local regions of the backgrounds, unsuccessful segmentations are obtained for the classes, such as *hot balloons* and *FC player*, as shown in Fig. 5.

The authors of [21] proposed a co-segmentation method by using linear anisotropic diffusion system. The method consists of two aspects. The first is single image segmentation method which segments the single image into K local regions by using linear anisotropic diffusion. The initial sources required for the linear anisotropic diffusion are automatically determined through clustering method. The second is common objects segmentation which jointly segments similar local regions by pushing the source placement of one image to be similar to its corresponding placement in other images. The model is finally formulated as K-way segmentation problem that maximizing the temperature on anisotropic heat diffusion. Since the energy function is submodular, the common local regions can be efficiently extracted by the greedy algorithm. In the experiment, the code released by the author is used.[2] To achieve accurate

[1]http://www.di.ens.fr/~joulin
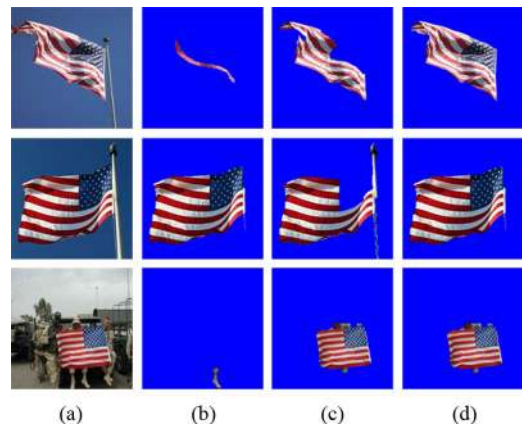
[2]http://www.cs.cmu.edu/~gunhee

Fig. 6. (a) Original images, where $flag$ is considered. (b)–(d) Results by considering color similarity, saliency maps, and the proposed method, respectively.

segmentation, we vary the scale of intra-image Gaussian weights. Furthermore, we adjust the number of segments (K) to obtain the co-segmentation. The co-segmentation results of the method in [21] are shown in the third row of each class block in Fig. 5. From the results, we can see that the common objects are accurately segmented from the image classes such as *flower2* and *bear*. Furthermore, the common objects are not successfully segmented from other image classes such as *goose2* and *hot balloons*, because it is difficult to locate seeded points of the common objects and backgrounds when the backgrounds are similar to each other.

The method in [20] selects interest common objects from local regions. In [20], a set of overlapping regions for each image are first extracted. Then a graph representing the similarity relationships between the regions is constructed. In the graph, the images are represented as nodes and regions for each image are labels of the node. The similarity is measured by random forest regressor which selects useful features from a total of 33 features. Finally, based on the graph, the segmentation problem is casted as a MRF setting problem which is solved by A* search algorithm. The main differences between the method of [20] and the proposed method are that first in the proposed method, the graph is constructed by considering neighboring images instead of all image pairs. Hence, low computational cost can be achieved by the proposed method. Secondly, in the proposed method, the local regions are represented by nodes and the images are represented by layers. The co-segmentation problem is casted as the shortest past problem. Thirdly, the proposed method uses the dynamic programming method to efficiently solve the co-segmentation problem. Fourthly, co-saliency model is employed to extract common objects from similar backgrounds in the proposed method. We implement the approach in [20] with Matlab codes. The parameters for the experiment are set according to [20]. The segmentation results of the method in [20] are shown in the fourth row of each class block in Fig. 5. From Fig. 5, it is seen that the method in [20] can achieve successful segmentations for the classes, such as *FC players*. Meanwhile, unsuccessful segmentations are observed for some images, such as *goose2*, because the segmentation results depend on the useful features learned from the training images. When the backgrounds are

Fig. 7. Co-segmentation results by considering shape. For each class, the first row shows original images. The following three rows display the experimental results of the methods in [16], [20], [21], and ours, respectively.

complex and similar to each other, selecting suitable training images for kinds of image groups is difficult.

In order to further illustrate our model, we use three approaches for calculating w_{ij} in (1). The first approach considers foreground similarity and uses w_{ij}^1 as w_{ij}. In the second approach, the saliency value is considered and $w_{ij} = w_{ij}^2$. The third approach is the proposed method that simultaneously considers the foreground similarity and the saliency value. The results of the three approaches are shown in Fig. 6, where three images of *flag* are shown. The original images are shown in Fig. 6(a). The results of the first approach, the second approach and the proposed method are shown in Fig. 6(b)–(d), respectively. It is seen that the proposed method achieves the best performance among these images. The mean error rate of all images in *flag* by the proposed method is 0.0757, which is lower than the first approach 0.1369 and second approach 0.1514.

3) Evaluation Using Shape Features: In shape feature based co-segmentation, we use the method in [25] to obtain shape description. Single image based saliency map obtained by [23] instead of co-saliency map is used to calculate $w_{ij,kl}^2$ in (1), because the proposed co-saliency map is to extract the co-saliency for multiple images based on color feature rather than shape feature. Furthermore, we do not treat the regions on the edge of the image as potential candidates, since they are similar to rec-

tangle. We delete these regions through setting the threshold of the ratio of the number of edge pixels in the local region to the total number of pixels on the edge. The segmentation results of shape based co-segmentation are shown in Fig. 7, where ten images of *applelogos* and *cow* are shown. For each class, the first row shows the original images. The segmentation results of the methods in [16], [21], and [20] are displayed in the second, third, and fourth rows, respectively. The results obtained by our proposed approach are shown in the last row. It is seen that the proposed method successfully extracts the common objects from the original images. For the other three methods, there are unsuccessful segmentations, such as the results of the last image of *applelogos*. The reason for the unsuccessful segmentation is that the methods proposed in [16] and [21] consider color feature instead of shape feature for the segmentation. For the method in [20], the common object segmentation is interfered by the other useful features.

Furthermore, we use the three approaches mentioned in Section III-A2 to calculate w_{ij} in (1). The results of the three approaches are demonstrated in Fig. 8, where the first row shows the original images. The results of the first, the second and the third approaches are displayed in the following three rows. The mean error rates of all the images in *applelogos* by the first, second, and third approaches are 0.0196, 0.2691,
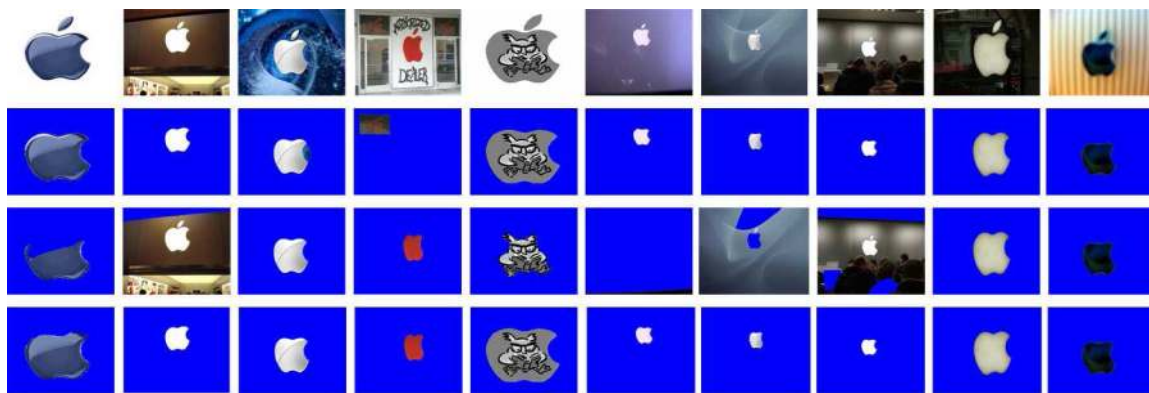
Fig. 8. First row: the original images. Second row: the experimental results by considering shape similarity. Third row: the experimental results by considering saliency maps. Last row: the experimental results by simultaneously taking shape similarity and saliency maps into account.

TABLE I
RESULTS COMPARISON BETWEEN THE PROPOSED CO-SEGMENTATION MODEL AND THE EXISTING METHODS
IN TERMS OF ERROR RATE. CLASSES IN MSRC AND CALTECH DATASETS ARE CONSIDERED

MSRC								
Method	Flowers1	Flowers2	Flowers3	Sheep	Cow1	Cow2	Airplane	Chair
[16]	0.3580	0.1817	0.2977	0.1729	0.3389	0.2710	0.3790	0.4643
[21]	0.0658	0.0312	0.1179	**0.0176**	0.0203	0.0309	0.1158	**0.0464**
[20]	0.0693	0.0273	0.1221	0.0475	0.0316	0.0547	**0.1001**	0.0863
Ours	**0.0088**	**0.0128**	**0.0433**	0.0237	**0.0192**	**0.0131**	0.1609	0.0958
Method	Doors	Trees	Spoons	Bicycles	Birds	Buildings	Cars	Chimneys
[16]	0.2101	0.4260	0.3512	0.3079	0.0558	0.4951	0.4326	0.1703
[21]	0.1414	0.2544	0.0699	**0.2514**	0.0529	**0.1665**	0.1406	0.1013
[20]	0.0574	0.1881	0.0421	0.3617	0.0362	0.2039	0.1737	0.1312
Ours	**0.0559**	**0.1577**	**0.0061**	0.3573	**0.0133**	0.2580	**0.1056**	**0.0252**
Method	Clouds	Leaves	Miscellaneous	Scenes	Signs	Windows	Doors	Cow
[16]	0.4107	0.3344	0.2032	0.1641	0.1156	0.2269	0.2101	0.1574
[21]	0.2472	0.4781	**0.1530**	0.3425	0.0342	0.1013	0.1414	0.1046
[20]	0.1825	**0.1853**	0.1674	0.4425	0.0567	0.1312	**0.0574**	0.0956
Ours	**0.1518**	0.2675	0.2962	**0.0659**	**0.0238**	**0.0788**	0.1115	**0.0199**
Caltech								
Method	Flag	Bear	Goose1	Grape	Bird	Face		
[16]	0.3584	0.4413	0.4954	0.4565	0.0991	0.4374		
[21]	0.1295	0.0636	0.1049	**0.1812**	**0.0919**	**0.1634**		
[20]	0.1219	0.1547	0.0812	0.2412	0.1664	0.2241		
Ours	**0.0757**	**0.0197**	**0.0481**	0.3274	0.1417	0.2432		

and 0.0078, respectively. It is seen that the proposed approach achieves the best performance among the three approaches.

B. Objective Evaluation

We employ objective criteria to evaluate our method. We evaluate the proposed co-segmentation method based on error rate, which is defined as the ratio of the number of wrongly segmented pixels to the total number of the pixels. The error rate is small when the object is accurately segmented. Since there are multiple images in a class, the average error rate over the error rates of all images is calculated to evaluate the performance. The error rates for the proposed method are shown in the last row of Tables I and II, where the results of the classes in MSRC and Caltech datasets are shown in Table I. Table II shows the results of classes in ICoseg and ETHZ datasets. (The classes segmented using the shape feature are *italic*.) The error rates by the proposed method are shown in the last row. The average percentage of the foreground across the image is 23.38% for the datasets used in the experiment. We can see from Tables I and II that our method achieves acceptable performance on most classes. Note that there are unsuccessful segmentations, such as

airplane and *grape*. The results for the unsuccessful segmentations are mainly caused by two reasons, i.e., the unsuccessful multiple local region generation and saliency object detection. For example, for the *grape*, the saliency maps are not successfully obtained from several images. Hence, the weights between common objects are affected, and the unexpected results are obtained. Furthermore, for *sheep* and *cow1* which have similar variability of the foreground, as the saliency detection method detects the shadows as saliency region for both two classes, the weights in the graph are interfered by the saliency detection. Hence, the error rates between *cow1* and *sheep* are slightly different with each other.

The error rates of [16], [21], and [20] are also shown in Tables I and II, where the second row shows the results of [16]. The results of [21] are shown in the third rows. The fourth row shows the results of [20]. The results show that the proposed method segments the common objects with the minimum error rates for most of the image pairs. Furthermore, we calculate the mean error rate by taking average over all image classes for comparison. The error rates of the methods in [16], [20], [21], and the proposed model are 0.2762, 0.1487, 0.1542, and 0.0946, respectively. It is seen that the error rate of the proposed

TABLE II
RESULTS COMPARISON BETWEEN THE PROPOSED CO-SEGMENTATION MODEL AND THE EXISTING METHODS IN TERMS
OF ERROR RATE. COLOR FEATURE IS CONSIDERED. CLASSES IN ICOSEG AND ETHZ DATASETS ARE CONSIDERED

IcoSeg

Method	Hot balloons	Goose2	FCplayer	Stonehenge	Ferrari	Christ	Helicopter	Gymnastics
[16]	0.3101	0.0983	0.1073	0.4780	0.3167	0.4123	0.0579	0.0968
[21]	0.1595	0.2306	0.0531	0.1684	0.0928	0.1919	0.1648	0.0698
[20]	0.0762	0.1338	0.0674	0.2309	0.1458	0.1684	0.0273	0.0411
Ours	**0.0559**	**0.0235**	**0.0379**	**0.1412**	**0.0852**	**0.1311**	**0.0088**	**0.0388**
Method	Bear	Sox	Taj	Pyramids	Elephants	Pandas	Airshows	Cheetah
[16]	0.3244	0.2562	0.4474	0.3509	0.3608	0.3299	0.0295	0.4722
[21]	0.1744	0.1562	0.1070	**0.1557**	0.1516	**0.2603**	0.1346	0.3218
[20]	0.1968	0.0392	0.1495	0.3949	0.1534	0.2617	0.0415	0.2049
Ours	**0.0921**	**0.0126**	**0.0449**	0.1870	**0.0940**	0.2942	**0.0046**	**0.1275**
Method	Kite1	Kite2	Skating	Soccer	Monks	Liberty	Speed	Track
[16]	0.0411	0.4156	0.1146	0.1409	0.3650	0.3094	**0.1066**	0.4061
[21]	0.2283	0.3531	0.0499	0.0672	0.1705	0.0789	0.1671	0.2043
[20]	0.1498	**0.1007**	0.0277	0.0762	0.0922	0.1847	0.4658	**0.1495**
Ours	**0.0164**	0.1671	**0.0073**	**0.0386**	**0.0475**	**0.0256**	0.1669	0.2874
Method	Windmill	Kendo	HOt balloons	Helicopter				
[16]	0.1762	0.1030	0.3101	0.0579				
[21]	0.2007	**0.0220**	0.1595	0.1648				
[20]	0.3796	0.1417	0.0762	0.0273				
Ours	**0.1098**	0.0633	**0.0147**	**0.0125**				

ETHZ

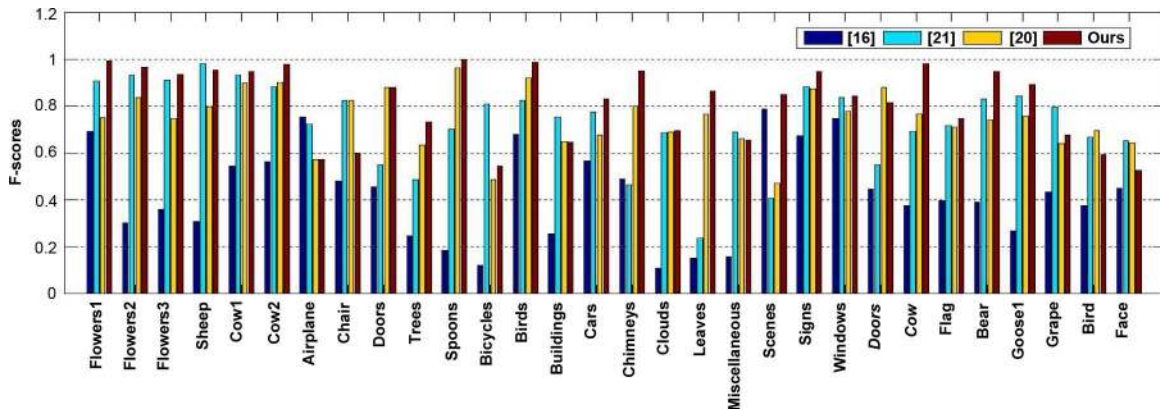| Method	*Applelogos*	*Bottles*	*Giraffes*	*Mugs*	*Swans*			
[16]	0.3264	0.2767	0.3272	0.4370	0.1533			
[21]	0.2191	0.1090	0.2236	0.2507	0.1440			
[20]	0.3356	0.2447	0.2066	0.2587	0.4227			
Ours	**0.0078**	**0.0917**	**0.0903**	**0.1384**	**0.0665**			

Fig. 9. Results comparison between the proposed co-segmentation model and the existing methods in terms of *F-scores*. Classes in MSRC and Caltech are considered.

method is about 0.0541 less than the method in [21] which obtains the minimum mean error rate among the other methods.

To completely verify the proposed co-segmentation method, another metric, i.e., *F-scores* in [27], is considered. The *F-scores* is calculated by $F\text{-}scores = 2Rec*Pre/(Rec+Pre)$. For a binary edge image, Precision (*Pre*) is the fraction of true contours among the contours of segmentation and Recall (*Rec*) is the fraction of ground-truth contours segmented. Note that an edge pixel is declared as true positive when the smallest distance between the pixel and the ground-truth pixels is less than a threshold (20 pixels). We use canny edge detection operator to obtain the edge image from the segmentations. The values of *F-scores* are shown in Figs. 9 and 10. The mean values of *F-scores* over all classes are 0.4335, 0.6438, 0.7233, and 0.8308 for the methods in [16], [20], [21], and the proposed method, respectively. We can see that the *F-scores* by the proposed co-segmentation method is larger than the other three methods.

We then compare our co-saliency model with the single image based saliency method [23]. We employ the co-saliency database given in [26] for evaluation, which is a challenging database and has 210 images with 105 classes. The precision recall curve, which is commonly used to evaluate the performance of saliency detection method, is used for performance measurement. In the experiment, we use both HC and RC proposed in [23] for calculating S_i. The precision and recall curves are shown in Fig. 11. The first two curves are obtained based on co-saliency database given in [26]. The rest two curves are obtained by our dataset. The first and third curves correspond to the method calculating S_i through HC. The second and the fourth curves are obtained by RC. As the plots show, our method improves the detection results of [23], since our curves are above the curves of [26]. The maximum *F-scores* for the four curves of the proposed method are 0.7945, 0.7054, 0.7568, and 0.7017, respectively. The maximum *F-scores* of the original method are 0.7402, 0.6846, 0.7261, and 0.6790, respectively. Compared

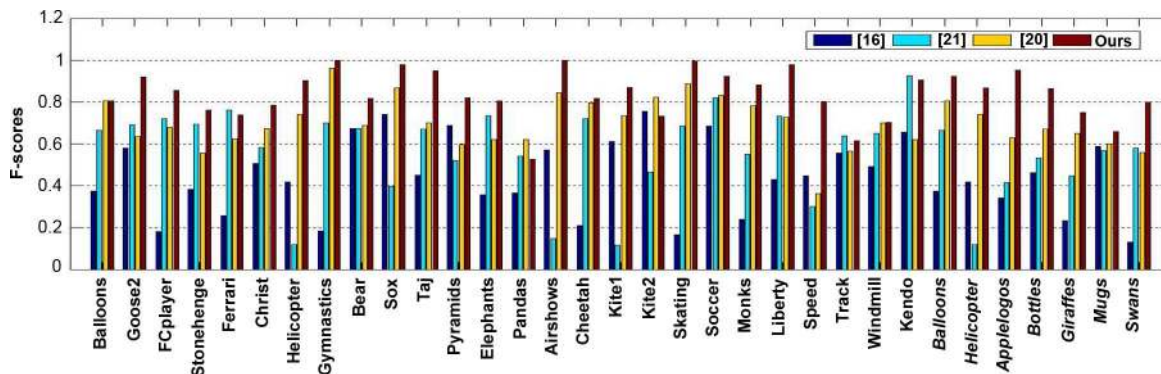Fig. 10. Results comparison between the proposed co-segmentation model and the existing methods in terms of *F-scores*. Classes in ICoseg and ETHZ are considered.
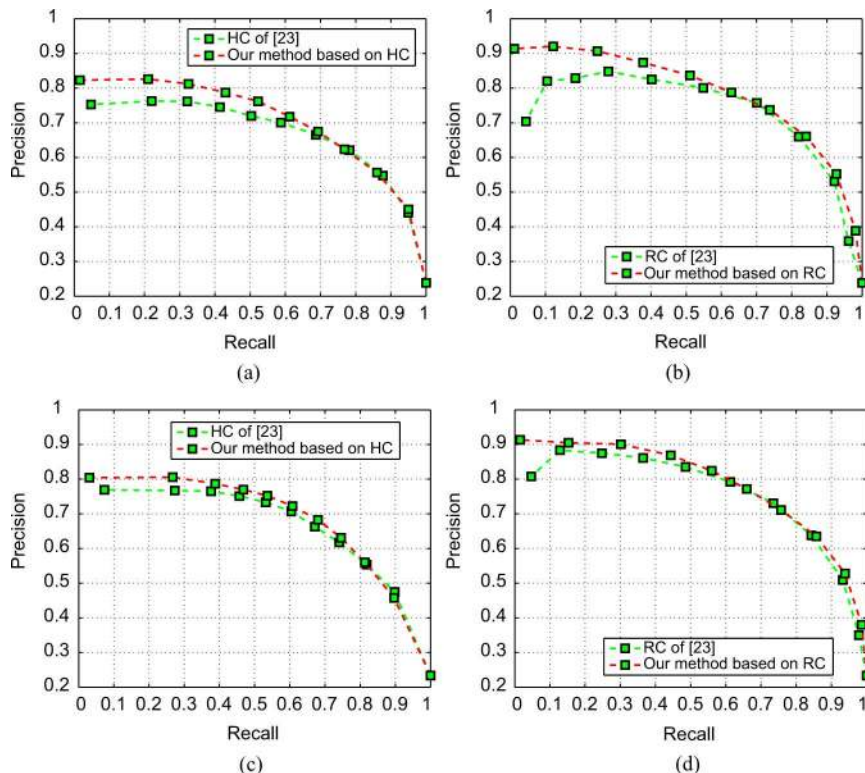

Fig. 11. Precision-recall curves for saliency detection by the methods [23] and our co-saliency model. (a) Curves for HC and our HC based method on the database given by [26]. (b) Curves for RC and our RC based method on the database given by [26]. (c) Curve for HC and our HC based method on the images used in this paper. (d) Curve for RC and our RC based method on the images used in this paper.

with original saliency detection method in [23], the proposed co-saliency model improves the saliency detection, since larger maximum *F-scores* are obtained.

C. Discussion

Since the digraph representing the similarity between the different local regions is constructed by the regions in the neighboring images, the change of image order can lead to different digraph for a group of images. When the common objects are segmented into local regions and the accurate saliency maps are obtained by saliency detection, the proposed method is robust to the change of the image order, since the path crossing the common objects always corresponds to the shortest path of the digraph. The feature similarities between the common objects and the saliency values of the common objects are large even

if the order of the images is changed. Fig. 12 shows the segmentation results with randomly ordered images. The segmentation results of the class *flower1* and *airplane* are displayed. For each class, the results of three random orders are displayed. The error rates for the six rows are 0.0088, 0.0088, 0.0088, 0.1376, 0.1133, and 0.1156, respectively. It is seen that the common objects are successfully segmented from the images with different orders.

Note that the video frames have an order. Hence, the proposed co-segmentation model is suitable for video segmentation. Fig. 13 presents the segmentation results, where the results of *diving* selected from video database given in [28] are shown. Fig. 13 shows that the proposed co-segmentation model can successfully segment the common objects from the video images. We also compare the proposed method with video segmentation method in [29] which segments each frame into a set of local regions. The similar regions in different frames are labeled with
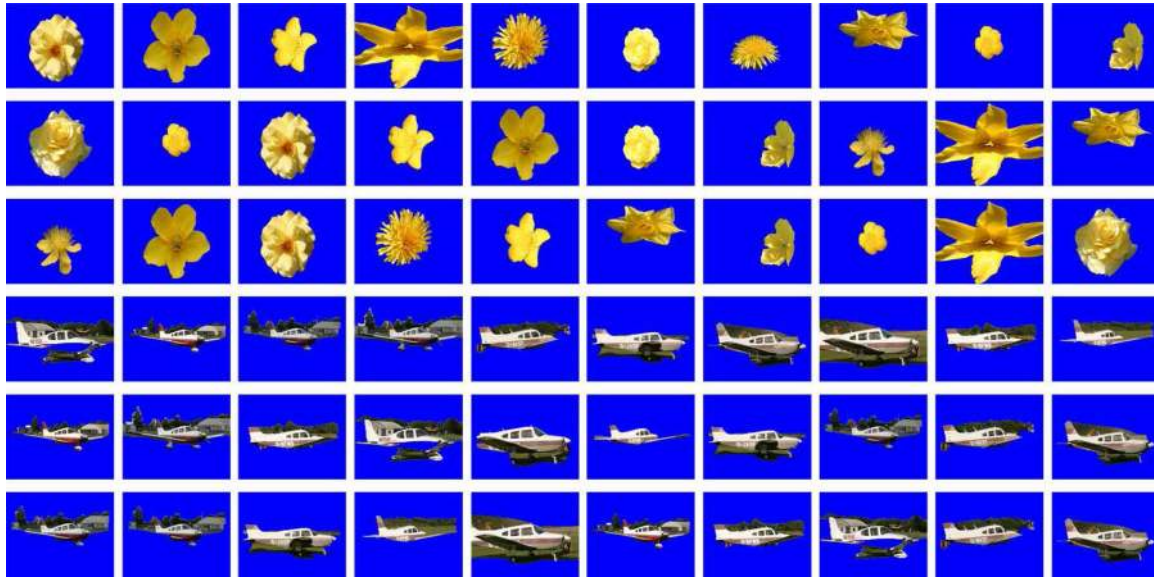
Fig. 12. Co-segmentation results with three randomly ordered images. The classes *flower1* and *airplane* are considered. The results in the three rows correspond to the results of the three random orders for each class.
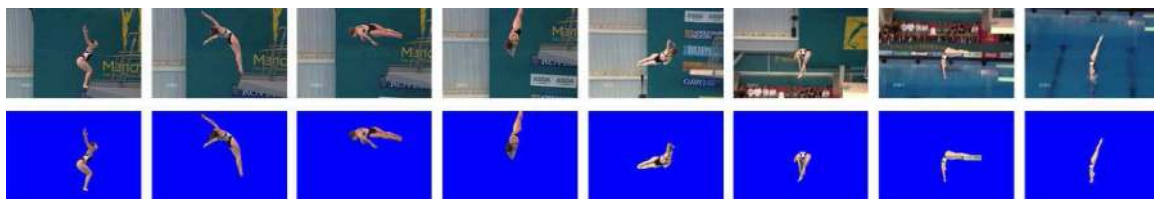

Fig. 13. Co-segmentation results for the video images. First row: original images. Second row: the co-segmentation results by the proposed method.

TABLE III

COMPARISON RESULTS BETWEEN OUR METHOD WITH VIDEO SEGMENTATION METHOD IN TERMS OF ERROR RATE

Method	diving	kicking	golf	lifting	riding	run	swing	walk
[29]	0.0070	**0.0166**	0.0173	**0.0176**	0.0109	0.0306	**0.0034**	0.0171
Ours	**0.0059**	0.0710	**0.0102**	0.0559	**0.0017**	**0.0070**	0.1631	**0.0066**

same number. We manually select the foreground regions in the first frames. The error rates of the method in [29] and the proposed method are shown in Table III. It is seen that the proposed method is comparable to the video segmentation method in [29].

For the proposed co-segmentation model, when there is more than one instance of the common object in an image, only one common object is extracted for each image as the shortest path contains only one node for each image. The examples are shown in Fig. 14, which shows that only one of common objects is extracted as the common objects. Note that when the common objects are adjacent, these common objects can be segmented since the multiple local region generation method can segment the combination of the common objects as a local region.

When the assumption that the common object is one of the local regions turns out to be false, the proposed co-segmentation method will extract part of the common objects as the local region. The reason is that the parts of the common objects are similar to each other, and the saliency values of these regions are large. In the experiment, we use superpixel segmentation method in [22] to obtain the local regions. We set $k = 100$ to guarantee that the common objects are generally segmented

Fig. 14. Co-segmentation results when there is more than one instance of the common object in an image. First row: original images. Second row: the results by the proposed method.

into several regions, which ensures that the assumption that the common object is one of the local regions is false. Fig. 15 shows that the common objects are partially segmented. For the image whose common object is not segmented as a local region, the segmentation result can be improved by the supervised segmentation method (such as Grabcut) using the segmentations of the other images. Note that different from the superpixel based segmentation which segments the image into many uniform and homogeneous regions with respect to the texture or color properties, we employ object detection and saliency detection to

Fig. 15. Co-segmentation results when the assumption that the common object is one of the local regions is false. First row: original images. Second row: the co-segmentation results.

achieve semantic segmentation. Hence, the common objects can be segmented as local regions with high probability by the proposed multiple local region generation model.

IV. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel co-segmentation model to segment common objects from multiple images. We first use multiple local region generation method to segment the original images into a variety of local regions. Then, based on the region similarities and saliency values, we construct a digraph to represent the relationships between different local regions. Finally, we formulate the co-segmentation problem as the shortest path problem, and we solve the problem using dynamic programming method. The experimental results demonstrate that the proposed model can efficiently segment the common objects from a group of images with lower error rate than many existing and conventional co-segmentation methods. In the future, we will extend the proposed co-saliency model by considering other features, such as texture and contour. Furthermore, we will design a method to generate the common template from the experimental results for a certain category.

REFERENCES

[1] H. LiK and N. Ngan, "Unsupervised video segmentation with low depth of field," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 12, pp. 1742–1751, Dec. 2007.

[2] N. Jacobson, Y.-L. Lee, V. Mahadevan, N. Vasconcelos, and T. Nguyen, "A novel approach to fruc using discriminant saliency and frame segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2924–2934, Nov. 2010.

[3] J. Han, K. N. Ngan, M. Li, and H. J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

[4] S. Goferman, L. Z. Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 2376–2383.

[5] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[6] F. Jing, M. Li, H. J. Zhang, and B. Zhang, "Relevance feedback in region-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 672–681, May 2004.

[7] H. Li, K. N. Ngan, and Q. Liu, "Faceseg: Automatic face segmentation for real-time video," *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 77–88, Jan. 2009.

[8] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," in *Proc. Int. Conf. Computer Vision*, vol. 1, pp. 105–112.

[9] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.

[10] J. Zhang, J. Zheng, and J. Cai, "A diffusion approach to seeded image segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 2125–2132.

[11] C. Rother, V. Kolmogorov, T. Minka, and A. Blake, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2006, pp. 993–1000.

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

[13] D. S. Hochbaum and V. Singh, "An efficient algorithm for cosegmentation," in *Proc. Int. Conf. Computer Vision*, Oct. 2009, pp. 269–276.

[14] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *Proc. Eur. Conf. Computer Vision*, 2010, pp. 465–479.

[15] D. Batra, D. Parikh, A. Kowdle, T. Chen, and J. Luo, "Seed image selection in interactive cosegmentation," in *Proc. IEEE Int. Conf. Image Processing*, Nov. 2009, pp. 2393–2396.

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

[17] D. Batra, A. Kowdle, and D. Parikh, "Icoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 3169–3176.

[18] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 1881–1888.

[19] K. Chang, T. Liu, and S. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 2129–2136.

[20] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 2217–2224.

[21] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. Int. Conf. Computer Vision*, Nov. 2011, pp. 169–176.

[22] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2009, pp. 2294–2301.

[23] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2011, pp. 409–416.

[24] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 73–80.

[25] H. Ling and D. Jacobs, "Shape classification using the inner-distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, Feb. 2007.

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

[27] B. Hariharan, P. A. ez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. Int. Conf. Computer Vision*, Nov. 2011, pp. 991–998.

[28] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. Eur. Conf. Computer Vision*, 2010, pp. 282–295.

[29] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 2141–2148.

Fanman Meng received the B.Sc. degree in computer science and technology and the M.Sc. degree in computer software and theory in 2006 and 2009, respectively. Since September 2009, he has been pursuing the Ph.D. degree in the Intelligent Visual Information Processing and Communication Laboratory (IVIPC) at University of Electronic Science and Technology of China (UESTC), Cheng Du, China.

His research interests include image segmentation, object detection, and visual attention.

Hongliang Li (SM'12) received the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, Xi'an, China, in 2005.

From 2005 to 2006, he joined the Visual Signal Processing and Communication Laboratory (VSPC) of the Chinese University of Hong Kong (CUHK) as a Research Associate. From 2006 to 2008, he was a Postdoctoral Fellow at the same laboratory in CUHK. He is currently a Professor in the School of Electronic Engineering, University of Electronic Science and Technology of China, Cheng Du, China. His research interests include image segmentation, object detection, image and video coding, visual attention, and multimedia communication system. he has authored or co-authored numerous technical articles in well-known international journals and conferences. He is a co-editor of a Springer book titled *Video Segmentation and Its Applications* (New York: Springer-Verlag, 2011).

Dr. Li was involved in many professional activities. He is a member of the Editorial Board of the *Journal on Visual Communications and Image Representation*. He served as TPC members in a number of international conferences, e.g., ISPACS2005, PCM2007, PCM2009, and VCIP2010, and served as Technical Program co-chair in ISPACS2009, and general co-chair of the 2010 International Symposium on Intelligent Signal Processing and Communication Systems. He will serve as a local chair of the 2014 IEEE International Conference on Multimedia and Expo (ICME). He was selected as the New Century Excellent Talents in University, Chinese Ministry of Education, China, in 2008.

Guanghui Liu received the M.Sc. and Ph.D. degrees in electronic engineering from University of Electronic Science and Technology of China (UESTC), Cheng Du, China, in 2002 and 2005, respectively.

In 2005, he joined Samsung Electronics, South Korea, as a senior engineer. Since 2009, he has been with the School of Electronics Engineering, UESTC, as an Associate Professor. His general research interests include digital signal processing and telecommunications, with emphasis on digital video transmission, and OFDM techniques. In these areas, he has published tens of papers in refereed journals or conferences, and received more than 10 patents (4 U.S. granted patents).

Dr. Liu served as the publication chair of the 2010 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2010).

King Ngi Ngan (F'00) received the Ph.D. degree in electrical engineering from the Loughborough University, Loughborough, U.K.

He is currently a Chair Professor at the Department of Electronic Engineering, Chinese University of Hong Kong. He was previously a Full Professor at the Nanyang Technological University, Singapore, and the University of Western Australia, Australia. He holds honorary and visiting professorships of numerous universities in China, Australia, and South East Asia. He has published extensively including 3 authored books, 6 edited volumes, over 300 refereed technical papers, and edited 9 special issues in journals. In addition, he holds 10 patents in the areas of image/video coding and communications.

Prof. Ngan served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, *Journal on Visual Communications and Image Representation*, *EURASIP Journal of Signal Processing: Image Communication*, and *Journal of Applied Signal Processing*. He chaired a number of prestigious international conferences on video signal processing and communications, and served on the advisory and technical committees of numerous professional organizations. He co-chaired the IEEE International Conference on Image Processing (ICIP) held in Hong Kong in September 2010. He is a Fellow of IET (U.K.) and IEAust (Australia), and an IEEE Distinguished Lecturer in 2006–2007.