A Novel Topic-level Random Walk Framework for Scene Image Co-Segmentation

Zehuan Yuan[1], Tong Lu[1*], and Palaiahnakote Shivakumara[2]

[1] National Key Laboratory of Software Novel Technology, Nanjing University, China
[2] Faculty of Computer Science and Information Technology, University of Malaya

Abstract. Image co-segmentation is popular with its ability to detour supervisory data by exploiting the common information in multiple images. In this paper, we aim at a more challenging branch called scene image co-segmentation, which jointly segments multiple images captured from the same scene into regions corresponding to their respective classes. We first put forward a novel representation named *Visual Relation Network* (VRN) to organize multiple segments, and then search for meaningful segments for every image through voting on the network. Scalable topic-level random walk is then used to solve the voting problem. Experiments on the benchmark MSRC-v2, the more difficult LabelMe and SUN datasets show the superiority over the state-of-the-art methods.

Keywords: image co-segmentation, voting, random walk, link analysis

1 Introduction

As one of the recent developments in computer vision, image co-segmentation has attracted the interest of researchers [10][7][16] [14][11][17][18][4][13][8] in the past years with its ability to remedy the loss of supervisory data by utilizing enhanced cues from co-occurring objects. However, although promising results have been achieved, most of them still face difficulties when dealing with scene images due to large intra-class variability and complex scene structures.

In this paper, we simultaneously analyze multiple images from the same scene and decompose each complex scene image into disjoint but meaningful segments with each corresponding to an instance of a scene object class (e.g., tree and car). We propose a fully automatic co-segmentation method that exploits both the appearance consistency of the same class and the spatial scene context constraints of different classes. The core of our method is to derive a directed *flowing-graph* named *Visual Relation Network* (VRN) (Sec.3) to characterize "soup of segments" [19] and their relations. In VRN, each node corresponds to an image segment and its latent class label is indicated by the state variable of the node. The statement of the *flowing-graph* means that the weight of any edge varies over the state variables of its linked nodes, like a valve controlling the water volume flowing from the starting point to the end. VRN thus succeeds in modeling both the appearance similarity and the spatial scene context relations between every two segments on class level by the form of adjustable weights. Note

that compared to bad segments, meaningful segments are believed to have strong intra-class appearance consistencies and spatial inter-class context relations with other segments. Thus they are actually the hubs of the graph with more water flowing into them. Thereby, co-segmentation from multiple scene images can be formulated as voting on a large-scale network. That is, by considering "topics" as "classes", we address co-segmentation by a topic-level random walk algorithm (Sec.4) on the *flowing-graph* to search for the meaningful segments that have high ranking scores. To achieve this, we use the greedy strategy (Sec.5) to search for the optimized segment combination from the selected meaningful segments. The overview of the entire framework is shown in Fig. 1. Note that since scene spatial context is unknown in advance, we thereby adopt a recursive way to alternate co-segmentation and learning stable spatial scene context (Sec.6).

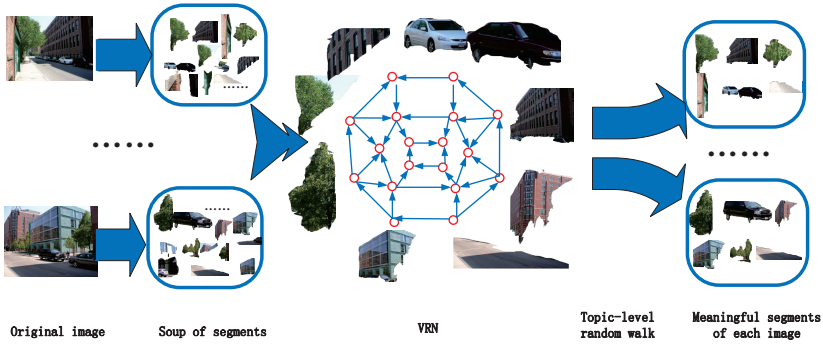

Fig. 1. The overview of the proposed method. The directed edges (blue arrow) in VRN model either the appearance similarity or the class-level spatial context relation between two segments (red circle).

Our main contributions include 1) the introduction of stable scene context into scene image co-segmentation, and 2) a new framework consisting of the VRN representation and the topic-level random walk on it to address the problem. Although topic-level random walk is familiar in mining social networks [26], it is novel for image co-segmentation to the best of our knowledge. According to the experiments on LabelMe [20] and SUN [1], we have averagely 10% improvement over the state-of-the-art methods. Moreover, the proposed VRN is sparse with few hubs [9] and thus is efficient for large scene datasets compared to popular pixel-label methods.

2 Related work

There are two branches towards image co-segmentation: two-class co-segmentation and multi-class co-segmentation. Two-class image co-segmentation aims to di-

vide every image into foreground or background regions with the former corresponding to objects. For comparison, multi-class image co-segmentation mainly focuses on the images consisting of many instances of different classes.

2.1 Two-class co-segmentation

The key step in these methods is to construct a proper appearance model to distinguish the two classes directly and robustly. Then it will be relatively easy to perform pixel-level labeling by using techniques like energy minimization. These methods are very different from each other in their selected features, such as color histogram [16], texture features [14], Garbos filters [6], stereo cues [11], objectness [24], and visual saliency [17].

To further propagate segmentation masks of common objects to different images, visually matching techniques across images have been introduced, typically consisting of region-level matching [18][4] and pixel-level correspondence [17]. [2] additionally models the shape of foreground objects explicitly by shape templates, thus getting better co-segmentation results by sharing shape templates among multiple foreground object instances. Recently, [25] further establishes consistent functional maps between two images in an reduced functional space to assist image co-segmentation. However, two-class co-segmentation algorithms can not be applied into the images that have many instances of different classes.

2.2 Multi-class image co-segmentation

As an extension to two-class co-segmentation, researchers have explored multi-class co-segmentation [10][7][13][8] recently. For example, [10] over-segments every image into multiple regions and labels each region under an combinatorial auction optimization framework. [7] converts the multi-class co-segmentation into the combination of spectral clustering and discriminative clustering, which well maintains the spatial structure of each image and the distinction among different classes. However, these methods only make use of the appearance consistency of one class across images. Actually, there also exists stable scene context across images that can be used for co-segmentation.

3 The proposed VRN

We introduce the construction of VRN in this section. Essentially, VRN is a weighted directed graph (V, E, W) with V as its vertex set, E as its edge set, and W as the edge weight set. Node a_i represents the i-th segment in the "soup of segments" of image a. For any image, we adopt the same strategy as [19] to obtain its "soup of segments", namely, we perform multiple rounds of graph-cut segmentation with a different parameter setting in each round. In addition, we add category-independent object proposals [3] into the soup to ensure meaningful segments of objects can be included into the soup. Note that we assume there is at least one meaningful segment in the soup for every class in the image.

A segment is described by the following two aspects: 1) appearance A that is characterized by pHOG, color distribution and texton distribution, and 2) class variable t belonging to $\{1, 2 \cdots, T\}$ and its distribution P that describes the probability of the segment belonging to one class in $\{1, 2 \cdots, T\}$. The unsupervised category discovery method [19] is used to initialize the distribution of every segment over different classes and cluster all the segments by Latent Dirichlet allocation (LDA). It encourages the segments in the same cluster to manifest similar appearance. Class label is initialized by $t = arg\max_c P(c)$. Edges will then be created between segments either from different images or inside the same image. See the example VRN in Fig. 2(a).

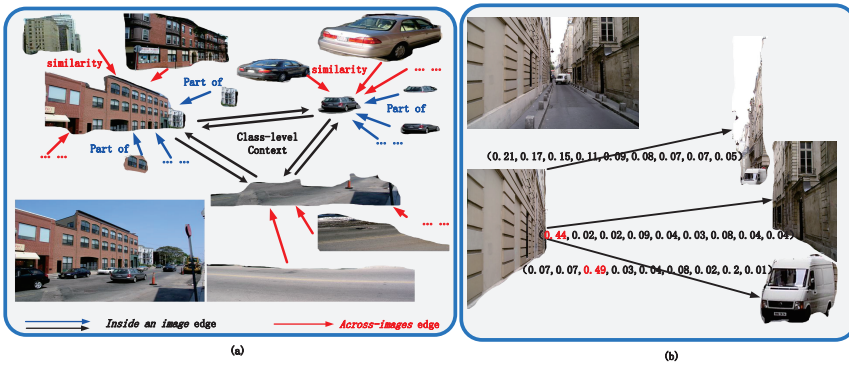

Fig. 2. An example VRN. (a) The VRN example, where three different object classes consisting of *building*, *sky* and *car* are denoted by class 1, 2 and 3, respectively. (b) Class-level weight vectors, in which the 9 elements respectively correspond to class pairs (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3). Note that for an edge between two pure segments, there is a significant peak in the weight vector compared to a flat distribution between mixed segments.

3.1 Edge construction across images

It is observed that meaningful segments tend to have consistent matches and thus we encourage the segments of similar appearance to be linked. Specifically, for each VRN node a_i, we first search for its K-nearest neighbors from "soups of segments" of other scene images by defining the following similarity measure S between two segments a_i and b_j:

$$S(a_i, b_j) = \frac{1}{|\ c\ |} \sum_c K_{\chi^2}(A_c(a_i), A_c(b_j)) \tag{1}$$

where A_c denotes the cth type of appearance features consisting of pHOG, texton and color, and $K_{\chi^2}(\cdot, \cdot)$ is a χ^2 kernel function. Generally, two visually similar

segments are more likely to belong to the same class, while the segments of different classes have dissimilar visual appearances. Note that no shape features are adopted to measure the similarity between two segments since the segment of an object may be a union of smaller ones due to over-segmentation or occlusion, and most importantly, many scene *stuffs* even have no explicit shapes.

Finally, let $\{b_j^k\}_{b\neq a}^{k=1,\cdots,K}$ denote the K-nearest neighbors of a_i, a *similarity* edge connecting a_i and b_j will be created, namely, the edge $a_i \rightarrow b_j$ as shown in Fig. 2(a) with its weight initialized as $S(a_i, b_j)$.

3.2 VRN construction inside an image

We hypothesize that two segments in one image should be connected if there exists a class-level spatial *context* relation or a *part-of* relation.

Establishing *part-of* edges As known, it is difficult to distinguish a part and the entire object without any semantic information. For instance, the building in Fig. 2(a) may be over-segmented and it is easy to consider any small segment as an independent meaningful object. Based on Gestalt Principles, we are prone to reserve the larger segments. Thereby, we define there exists a *part-of* relation between two segments if one segment belongs to another one. For this case we add an edge to link them. That is, given two segments a_i and a_j with an overlapped scale $\frac{a_i \cap a_j}{\min(a_i, a_j)} > 0.95$, we add a directed edge (a_i, a_j) if the segment a_i is smaller, otherwise (a_j, a_i) is added. The weight of a *part-of* edge is fixed as $\tau = 0.52$.

Inter-class spatial context The inter-class spatial context represents that the instances of different classes in a scene have a roughly stable spatial layout. We take a street scene as an example and assume there exist instances of class 1, class 2 and class 3 which correspond to "pedestrian", "road" and "sky", respectively. The instances of class 1 always walk on the instances of class 2, and similarly the instances of class 3 will be above on all the instances of the rest two classes in this scene. However, due to perspective deformation, 3D spatial context of a scene manifests in a diverse way in 2D image space. In our method, we classify 2D inter-class spatial context into different clusters. In another word, the images that have similar spatial structures are grouped into the same cluster, and thus the inter-class spatial context for each cluster are consistent and stable. Each image has a cluster label that corresponds to the group it exists. Specifically, we first extract the global Gist feature [15] to represent its spatial structure for every image. Then the hierarchical agglomerative clustering is used since it chooses the number of clusters in an automatical way. Fig. 3 shows a tree view example of clusters, where the images that have visually similar elements and layout are categorized into the same cluster.

For each cluster c, we use image-dependent location probability maps [5] to model its inter-class spatial context. A map $M^c_{t_j|t_i}(\cdot,\cdot)$ corresponds to the class

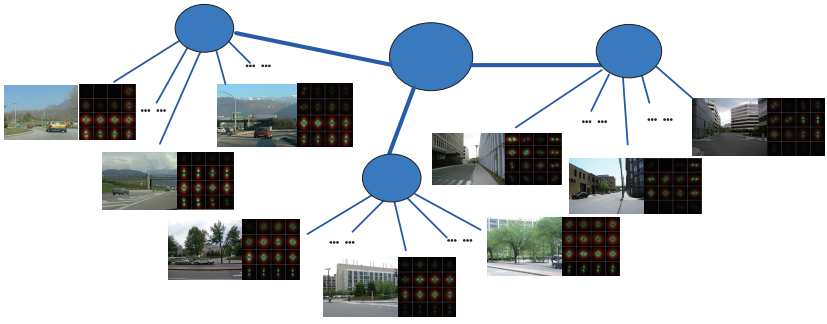

Fig. 3. A hierarchical tree view example after clustering. Example images and their corresponding Gist features are shown for each cluster.

pair (t_j, t_i) that models the preference of t_j at any relative location to pixel of t_i. For example, $M^c_{t_j|t_i}(x, y)$ encodes the probability of the pixel with (x, y) deviation to anyone pixel of t_i belonging to t_j.

Establishing *context* edges Based on the learned inter-class spatial context model, *context* edges are added to include the inter-class spatial constraints between two segments in an image. Note that the strength of an edge is a function over the class variables of its linked segments. Specifically, given two segments a_i and a_j with the overlapped scale $\frac{a_i \cap a_j}{a_i \cup a_j} < 0.05$, we add two directed edges (a_i, a_j) and (a_j, a_i) into E. With each element corresponding to a class pair, the weight of (a_i, a_j) is a class-level vector $c(t_i, t_j)$ indicating the strength of spatial relation if a_i and a_j are equal to t_i and t_j, respectively:

$$c(t_i, t_j) = \frac{p(t_i|a_i)}{\mid a_i \mid\mid a_j \mid} \sum_{\substack{(x,y)\in a_i \\ (x^*,y^*)\in a_j}} M^c_{t_j|t_i}(x^* - x, y^* - y) \tag{2}$$

where $\mid a_i \mid$ and $\mid a_j \mid$ are the number of pixels in a_i and a_j, respectively. $p(t_i|a_i)$ is the probability of a_i under class t_i, and $M^c_{t_j|t_i}(\cdot, \cdot)$ is the relative location map of t_j given t_i in the cluster c of a. Note that the weight of (a_j, a_i) may not be the same as that of (a_i, a_j) because of different relative location maps. Note that for an edge between two pure segments, there is a significant peak in the weight vector compared to a flat distribution between mixed segments (See Fig. 2(b)). This can be utilized by our topic-level random walk later to search for meaningful segments.

4 Topic-level random walk on VRN

After initializing all the nodes and the edges in VRN, it is observed that meaningful segments perform the role of hubs to which many other nodes are directed. Thus we consider the weight of every directed edge as a vote from the starting segment to the ending one. Since the weight of a *context* edge is a function of class variables, we adopt a topic-level random walk method to improve the ranking quality by integrating class-level spatial context.

Specifically, given a node a_i, we introduce a ranking score vector $\{r[a_i,t]_{t=1,\cdots,T}\}$ to represent the importance of a_i under class t, rather than a simple important value used in Pagerank. Votes are then derived from either the linked segments in the same image a or those from other scene images b. The vote of a *similarity* edge is essentially intra-class. That is, for an edge (a_i,b_j), a_i only votes $r[a_i,t]$ to $r[b_j,t]$. However, for a *context* edge (a_i,a_j), the vote is inter-class. If a_i and a_j respectively have strong spatial relations under classes t_i and t_j, namely $c(t_i,t_j)$ is large, a_i votes $r[a_i,t_i]$ to $r[b_j,t_j]$. This encourages meaningful segments to have a higher ranking score under its correct class label. The vote of a *part-of* edge is not class-level because we prefer to the larger regions regardless of their classes. Mathematically, the topic-level ranking score of a_i under class t can be recursively defined by:

$$r(a_i,t)=\varepsilon\frac{p(t|a_i)}{\mid V\mid}+(1-\varepsilon)\Big(\kappa\sum_{(b_j,a_i)\in E} r(b_j,t)w_{b_ja_i}+ (1-\kappa)\sum_{(a_j,a_i)\in E} V(a_j,a_i,t)\Big)$$

$$V(a_j,a_i,t)=\begin{cases}\sum_{t_j}\tau r(a_j,t_j) & (a_j,a_i)\text{is } \textit{Part of}\\ \sum_{t_j} c_{a_ja_i}(t_j,t)r(a_j,t_j) & \text{Otherwise}\end{cases} \tag{3}$$

where ε is the damping factor and is set by a typical value 0.15. κ represents the balance factor between two types of edges. $w_{b_ja_i}$ is the normalized appearance similarity measure $S(a_i,b_j)$ and E represents all the edges in VRN. Intuitively, a VRN node with a relatively high ranking score is much likely to connect with the VRN nodes that also have high ranking scores.

The proposed topic-level random walk can be further reduced into a simple Pagerank representation. The details are included in the supplementary material. Accordingly, the iterative definition of topic-level random walk is promised to converge theoretically. The inference of $r[\cdot,t]$ of any node in VRN can thus be addressed by the Power method used for Pagerank.

5 Segments selection and class inference

A meaningful segment in general has a relatively high ranking score. To search for meaningful segments from any scene image a, we adopt a greedy algorithm and the details are shown in Tab. 1. Since every segment has an importance

vector, we calculate its overall importance. One segment with a high overall importance is considered more important. Note that we infer the class label for any selected segment a_i by $\bar{t} = \arg\max r(a_i, t)$.

Table 1. The greedy algorithm to choose meaningful segments.

input: Image set D and all the candidate segments S **Output:** The selected segments for every image in D
For any image a in D **(1)** Calculate the overall score $r_{overall}$ of a segment a_i and assign a class label $\bar{t}$ to it by $\bar{t} = \arg\max r(a_i, t), m_a = \frac{1}{\|T\|}\sum_t r(a_i, t), v_a = \frac{1}{\|T\|}\sum_t (r(a_i, t) - m_a)^2$, $r_{overall} = v_a * \max r(a_i, t)$; **(2)** Sort all $\{a_i\}_{i=1,\cdots,S_a}$ by $r_{overall}$ and initialize the selected segments set $segC = []$; **(3)** Select a_i in the descending order of $r_{overall}$; **(4)** For $a_j \in segC$ calculate $Overlap = \frac{a_i \cap a_j}{a_i \cup a_j}$, if $Overlap > 0.1$ return **(3)**; **(5)** Add a_i into $segC$, if $\bigcup segC < 0.9 *$ imagesize of a, return **(3)**. **End**

6 Iterative scene image co-segmentation using VRN

After constructing VRN, we adopt an iterative strategy to perform co-segmentation and update scene spatial context. It will converge to an optimal solution when scene spatial context are stable. The overall framework is as follows:

1. **Initialization-step:** Initialize the VRN representation as introduced in Sec. 3.
2. **Iteration-step:** Search for meaningful segments for every image iteratively:
   (a) Use topic-level random walk on the VRN to calculate ranking scores of all the nodes in it;
   (b) Select meaningful segments and infer their class variables in every scene image;
   (c) Calculate a new VRN representation by updating the inter-class spatial context for every cluster and the class distribution associated to each node.

In this stage, the *context* edges should only be recalculated based on the new inter-class context model. Therefore, we first update inter-class spatial context for each scene cluster according to the selected segments and their class labels. That is, we need update a lot of relative location maps by recalculating $M^c_{t2|t1}$ between any two classes of $t2$ and $t1$ for each cluster c. Note that only the images with the cluster label c are used to update $M^c_{(\cdot,\cdot)}$. Specifically, given a pixel p_1 of the class $t1$, $M^c_{t2|t1}(u, v)$ counts the ratio of pixels p_2 at the offset (u, v) to any p_1

of the class $t1$ belonging to $t2$. The map $M^c_{t2|t1}(u,v)$ is maintained in normalized image coordinates $(u,v) \in [-1,1] \times [-1,1]$. We also have $\sum_{t2} M^c_{t2|t1}(u,v) = 1$ so that $M^c_{t2|t1}$ represents a proper conditional probability distribution over the class $t1$. See details in [5].

Next, for a segment a_i in V of VRN, we update its class distribution by

$$p(t_i|a_i) = \frac{r(a_i,t_i)}{\sum_t r(a_i,t)} \tag{4}$$

Accordingly, we have a new class distribution for every segment to construct a new VRN as in Sec. 3.

7 Experiments and discussions

7.1 Experimental settings

We employ the normalized cuts algorithm [21] to generate the "soup of segments" of every image in a scene. Specifically, we vary the segment number from 3 to 12 and accordingly run the algorithm 10 times. The top 10 object proposals are also added into the soup using [3]. Thus totally 85 segment candidates are obtained for every scene image in our dataset. We then use the algorithm [19] to initialize the class distribution for each candidate segment. The appearance characteristic A of each segment includes three types of Bag-of-features histograms: Texton Histograms (TH), Color Histograms (CH), and pyramid of HOG (pHOG). We generate these histograms in the same way as [12]. κ is a weight to balance the importance of context and appearance in topic-level random walk, which is fixed as $\kappa = 0.65$ because we find κ and τ are unsensitive to specific scenes as long as the scenes have stable context. Thereby, we get their respective optimal values by a simple validation set.

7.2 Datasets

We evaluate our method on three datasets: MSRC-v2 [22], LabelMe [20] and SUN [1]. MSRC-v2 has altogether 21-classes (591 images). We pick up the images that have more than 3 classes for testing and thus form a subset (380 images) consisting of 13 classes. The images from LabelMe and SUN are collected from realistic daily life scenes. We choose six scenes: office (180 images) (LabelMe), movie theater (32 images) (LabelMe), bathroom (350 images) (LabelMe) and bedroom (307 images) (LabelMe), static street scene (400 images) (SUN) and outdoor (137 images) (SUN), with each scene category consisting of more than five object classes. We normalize all the images from LabelMe and SUN into 256×256 to avoid scale variations. Note that all the images in our dataset have pixel-level ground-truth labels.

7.3 Evaluation

Firstly, we adopt the segmentation accuracy to quantitatively evaluate our results. For each class, we denote the ground-truth segments and the obtained segments with G and C , respectively. Then the segmentation accuracy can be defined as the ratio of the intersection of G and C to the union of them, namely $\frac{G \cap C}{G \cup C}$. Besides, *purity* score [23] is also adopted to measure the coherency of class labeling of our method over the entire dataset. For each selected segment, its ground-truth class label is the one that the majority of pixels in it belong to. Note that different class labels may be potentially assigned to the selected segments with the same ground-truth class label in different images.

7.4 Scene co-segmentation results

Segmentation accuracy on MSRC-v2 For each image, we search for meaningful scene segments from its 85 segment candidates. Eight examples of four scenes in our subset of MSRC-v2 are illustrated in Fig. 4. It can be seen that most classes in these images are well segmented.

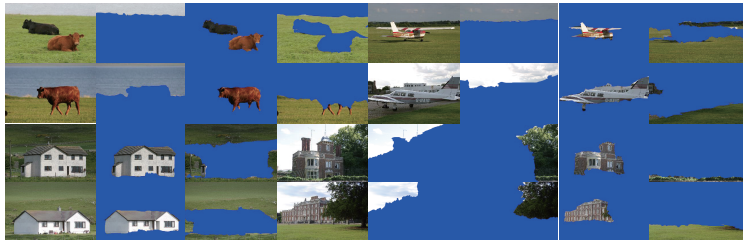

Fig. 4. Image co-segmentation examples of MSRC-V2. Two images are shown for each scene. The right images are the union of the selected meaningful segments of the same class from the original images.

The segmentation accuracies for MSRC-v2 are listed in Tab. 2. Firstly, we compare our complete version with the modified versions to test the effectiveness of the spatial context (see (b) in Tab. 2) and the *part-of* relation (see (c)) on MSRC-v2. Additionally, an appearance-only approach for image co-segmentation without constructing the *inside the same image* edges of VRN is also performed for comparisons (see (d)). It can be seen that the complete version of the proposed approach performs best. Thereby, the *part-of* relation helps avoid oversegmenting scene elements into smaller parts. Moreover, the spatial context and the *part-of* relations can supplement with each other to improve the performance.

We further compare our approach with the baseline algorithm [19], the recent unsupervised object discovery method [12] and another two state-of-the-art

multi-class co-segmentation approaches Jour [7] and Kim [10]. Although the baseline [19] and [12] aim at category discovery, they also output segments of each category. Thus comparisons are available using their public codes. Note that we do not compare with [12] directly due to their priors of known scene elements. We adopt their version without object-graph. Similarly, we use public codes of [7] and [10], and then adjust parameters to get their best results. We can see our method performs best in 6/13 classes and obtains competing results over the others. The main reason is that most images consisting the rest 7 classes have few objects and large inter-class variability. Thus the methods based on only appearance is sufficient to discriminate them. Note that [10] performs relatively bad on MSRC-v2 due to their strong assumption for multi-class co-segmentation. Thereby, the results validate that the spatial context and *part-of* relation play a critical role in selecting meaningful segments.

Table 2. Accuracy comparisons on MSRC-v2.

Class	Propose	(b)Cont	(c)Part	(d)Appr	Russ[19]	Lee.[12]	Jour[7]	Kim.[10]
car	0.51	0.45	0.42	0.40	0.31	0.38	**0.57**	0.44
sky	**0.81**	0.75	0.79	0.73	0.67	0.75	0.80	0.52
Tree	0.57	0.51	0.54	0.47	0.57	0.48	**0.61**	0.49
Grass	0.56	0.56	0.55	0.51	0.49	0.53	**0.57**	0.50
Building	**0.63**	0.54	0.56	0.50	0.45	0.49	0.51	0.51
House	**0.67**	0.55	0.56	0.50	0.54	0.57	0.52	0.44
Road	**0.61**	0.59	0.58	0.57	0.41	0.44	0.60	0.51
Cow	0.53	0.46	0.51	0.51	0.40	**0.54**	**0.54**	0.49
Plane	**0.49**	0.47	0.42	0.40	0.38	0.44	0.45	0.31
Sheep	0.62	0.55	0.59	0.60	0.47	0.63	0.66	**0.68**
Bird	0.46	0.44	0.45	0.41	0.34	0.40	**0.47**	**0.47**
Dog	0.42	0.37	0.38	0.35	0.39	0.35	0.41	**0.47**
Boat	**0.38**	0.43	0.40	0.39	0.38	0.32	**0.38**	0.34

Segmentation accuracy on LabelMe and SUN One example image and its segmentation results of every scene in LabelMe and SUN are shown in Fig. 5. Although there are many object classes in these images, our method can well discriminate them and successfully select meaningful segments.

For the scene images from LabelMe and SUN, we average the segmentation accuracy of all the classes in each scene because of the large amounts of categories in them. The results are illustrated in Tab. 3. We find the average accuracy of MSRC-v2 is higher than those of LabelMe and SUN. This is due to the fact that most of the scenes in MSRC-v2 have fewer categories and large inter-class variability. From the comparison results, we can see our methods perform overwhelmingly better that other methods. It follows our intuition that there exists stable scene context in each scene and they can help much to discriminate

Fig. 5. Scene image co-segmentation examples on LabelMe and SUN. From top to bottom: Outdoor, Bathroom, Bedroom, Movie Theater, Static Office and Static Street. The first column represents an example image, while the rest columns are the results after co-segmentation. The selected segments are ranked in a decreasing order by their overall importance scores from left to right. The blue subgraphs at the tail of each row are only for alignment.

Table 3. Accuracy comparisons on LabelMe and SUN.

Scene	Propose	(b)Cont	(c)Part	(d)Appr	Russ[19]	Lee.[12]	Jour[7]	Kim.[10]
Office	**0.35**	0.29	0.22	0.20	0.25	0.30	0.29	0.24
Theater	**0.44**	0.40	0.34	0.29	0.31	0.34	0.35	0.31
Bathroom	**0.45**	0.39	0.38	0.39	0.35	0.32	0.33	0.39
Bedroom	**0.39**	0.32	0.30	0.25	0.30	0.29	0.30	0.38
Street	**0.52**	0.42	0.40	0.39	0.39	0.45	0.41	0.40
Indoor	**0.44**	0.35	0.32	0.30	0.36	0.30	0.37	0.33

different classes. To conclude, our method succeeds in combining appearance, *part of* relation and the scene context to select meaningful segments.

Purity on MSRC-v2, LabelMe and SUN *Purity* scores of two datasets are illustrated in Tab. 4. We find that our method succeeds in assigning consistent class labels to each scene element category. Averagely, the *purity* scores on LabelMe and SUN are lower than that on MSRC-v2 due to large intra-class variability. Since Jourlin [7]and Kim [10] perform pixel labeling, we calculate their *purity* scores at pixel-level. Overall, it is consistent with segmentation accuracy results and thus our method is better than other methods. To conclude, our method can not only select meaningful segments for any image but also assign consistent class labels to these segments of the same category.

Table 4. Purity on MSRC-v2, LabelMe and SUN.

Dataset	Propose	(b)Cont	(c)Part	(d)Appr	Russ[19]	Lee.[12]	Jour[7]	Kim.[10]
MSRC-v2	0.79	0.63	0.64	0.50	0.51	0.77	**0.80**	0.77
LabelMe	**0.52**	0.45	0.40	0.28	0.34	0.46	0.46	0.41
SUN	**0.58**	0.44	0.45	0.40	0.37	0.47	0.45	0.47

7.5 Impacts of class number

The class number T in our aforementioned experiments is fixed to achieve the best performance. In this section, we evaluate the influence of different T against *purity*. From the results on three datasets (the front two images in Fig. 6), we can see that the performance reaches the best when selecting the class number as 1.4× *Number of scene element categories*.

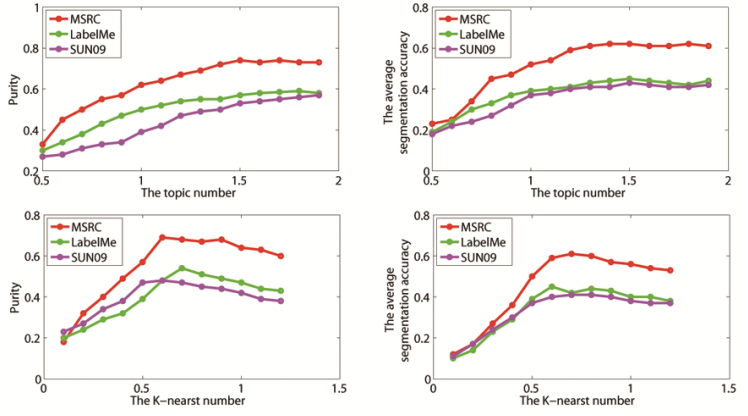

Fig. 6. The first two graphs show the performances as the class number varies, while the last two graphs correspond to the sparsity variations of VRN. In the temporary context, the class number and K-nearest number of the horizontal axis is a percentage to the amount of scene element categories and scene image number, respectively.

7.6 Impacts of the sparsity measurement of VRN

The sparsity of the VRN is controlled by K when constructing *across images* edges. Generally, a large K can cause the complexity of topic-level random walk, while a small K is insufficient to find enough matches during image co-segmentation. The last two images of Fig. 6 show that the range [0.6, 0.8] is the best choice. Moreover, too many matches can cause a negative impact. The main reason is that too many extra intra-class votes will mislead topic-level random walk to derive error importance scores.

7.7 Running time

We implement our entire algorithm by Matlab and run on our machine with Intel i3-2130 CPU@ 3.40GHz. When "soups of segments" are available, the construction of VRN and the selection of meaningful segments are relatively quick. Without any optimization of codes, it takes about 10 mins to construct the overall VRN and 2 mins to select meaningful segments for each image. The overall co-segmentation requires 44 min for convergence for 400 images comprising totally 34000 segments. Compared to pixel-label methods, it is a valuable step that benefits from our link analysis extension to Pagerank.

8 Conclusion

In this paper, we present a novel *visual relation network* to model the relationship between scene segment candidates and perform topic-level random walk on the network to exploit scene co-segmentation. The experiments on different datasets show the effectiveness of our method. However, if unfortunately most of the candidate segments are "garbage" ones, the accuracy will be according decreased during image co-segmentation. Potentially it can be avoided by enriching "soup of segments". Our further work is to improve the accuracy of our unsupervised scene image co-segmentation by including more class-level context cues.

Acknowledgment

The work described in this paper was supported by the Natural Science Foundation of China under Grant No. 61272218 and No. 61321491, the 973 Program of China under Grant No. 2010CB327903, and the Program for New Century Excellent Talents under NCET-11-0232.

References

1. Choi, M.J., Torralba, A., Willsky, A.S.: A tree-based context model for object recognition. IEEE Trans. Pattern Anal. Mach. Intell. 34(2), 240–252 (2012)
2. Dai, J., Wu, Y.N., Zhou, J., Zhu, S.C.: Cosegmentation and cosketch by unsupervised learning. In: ICCV (2013)
3. Endres, I., Hoiem, D.: Category-independent object proposals with diverse ranking. IEEE Trans. Pattern Anal. Mach. Intell. 36(2), 222–234 (2014)
4. Faktor, A., Irani, M.: Co-segmentation by composition. In: ICCV (2013)
5. Gould, S., Rodgers, J., Cohen, D., Elidan, G., Koller, D.: Multi-class segmentation with relative location prior. International Journal of Computer Vision 80(3), 300–316 (2008)
6. Hochbaum, D.S., Singh, V.: An efficient algorithm for co-segmentation. In: ICCV. pp. 269–276 (2009)
7. Joulin, A., Bach, F., Ponce, J.: Multi-class cosegmentation. In: CVPR. pp. 542–549 (2012)

8. Joulin, A., Bach, F.R., Ponce, J.: Discriminative clustering for image co-segmentation. In: CVPR. pp. 1943–1950 (2010)
9. Kim, G., Faloutsos, C., Hebert, M.: Unsupervised modeling of object categories using link analysis techniques. In: CVPR (2008)
10. Kim, G., Xing, E.P.: On multiple foreground cosegmentation. In: CVPR. pp. 837–844 (2012)
11. Kowdle, A., Sinha, S.N., Szeliski, R.: Multiple view object cosegmentation using appearance and stereo cues. In: ECCV (5). pp. 789–803 (2012)
12. Lee, Y.J., Grauman, K.: Object-graphs for context-aware visual category discovery. IEEE Trans. Pattern Anal. Mach. Intell. 34(2), 346–358 (2012)
13. Ma, T., Latecki, L.J.: Graph transduction learning with connectivity constraints with application to multiple foreground cosegmentation. In: CVPR. pp. 1955–1962 (2013)
14. Mukherjee, L., Singh, V., Peng, J.: Scale invariant cosegmentation for image groups. In: CVPR. pp. 1881–1888 (2011)
15. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
16. Rother, C., Minka, T.P., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In: CVPR (1). pp. 993–1000 (2006)
17. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: CVPR. pp. 1939–1946 (2013)
18. Rubio, J.C., Serrat, J., López, A.M., Paragios, N.: Unsupervised co-segmentation through region matching. In: CVPR. pp. 749–756 (2012)
19. Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR (2). pp. 1605–1614 (2006)
20. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. International Journal of Computer Vision 77(1-3), 157–173 (2008)
21. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
22. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV (1). pp. 1–15 (2006)
23. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley (2005)
24. Vicente, S., Rother, C., Kolmogorov, V.: Object cosegmentation. In: CVPR. pp. 2217–2224 (2011)
25. Wang, F., Huang, Q., Guibas, L.J.: Image co-segmentation via consistent functional maps. In: ICCV (2013)
26. Yang, Z., Tang, J., Zhang, J., Li, J., Gao, B.: Topic-level random walk through probabilistic model. In: APWeb/WAIM. pp. 162–173 (2009)