Unsupervised Cosegmentation based on Global Graph Matching

Takanori Tamanaha
Grad. School of IST
The University of Tokyo
Tokyo, Japan
tamanaha@nlab.ci.i.u-tokyo.ac.jp

Hideki Nakayama
Grad. School of IST
The University of Tokyo
Tokyo, Japan
nakayama@ci.i.u-tokyo.ac.jp

ABSTRACT

Cosegmentation is defined as the task of segmenting a common object from multiple images. Hitherto, graph matching has been known as a promising approach because of its flexibility in matching deformable objects and regions, and several methods based on this approach have been proposed. However, candidate foregrounds obtained by a local matching algorithm in previous methods tend to include false-positive areas, particularly when visually similar backgrounds (e.g., sky) commonly appear across images.

We propose an unsupervised cosegmentation method based on a global graph matching algorithm. Rather than using a local matching algorithm that finds a small common subgraph, we employ global matching that can find a one-to-one mapping for every vertex between input graphs such that we can remove negative regions estimated as background. Experimental results obtained using the iCoseg and MSRC datasets demonstrate that the accuracy of the proposed method is higher than that of previous graph-based methods.

Categories and Subject Descriptors

I.4.6 [Image Processing and Computer Vision]: Segmentation - *regiongrowing*, *partitioning*

General Terms

Algorithms

Keywords

Cosegmentation, Global Graph Matching, GrabCut

1. INTRODUCTION

Constructing a visual knowledge base for image recognition from Web data has been studied actively in the field of computer vision [1, 2]. While images on the Web relevant to a specific visual concept are usually collected via text-based image search engines, they often include miscellaneous backgrounds that hamper the construction of a visual knowledge base. Unsupervised cosegmentation is the technique of segmenting a common foreground (object region) from multiple images and can be used for removing backgrounds irrelevant to each visual concept to construct a high-quality knowledge base.

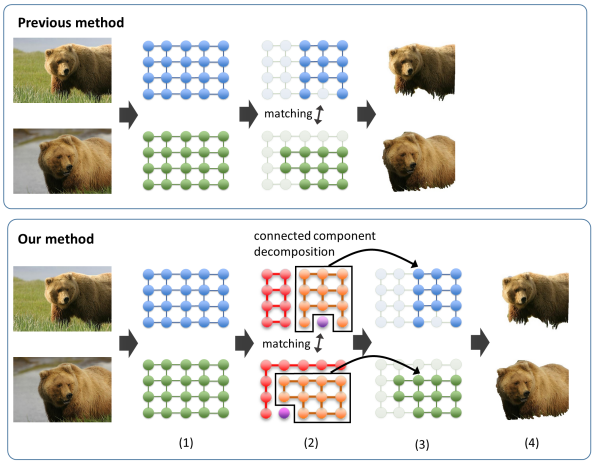

Figure 1: Process flows of previous typical methods (top) and the proposed method (bottom). Given two images, (1) each image is transformed into a graph composed of superpixels. (2) Visually similar regions (subgraphs) are paired by global graph matching and (3) the subgraph representing the foreground is extracted by connected component decomposition. Finally, (4) the foreground is refined by GrabCut using the subgraph as a seed.

Cosegmentation was first introduced by Rother et al. [3], and a number of methods have been proposed since then. Vicente et al. [4] proposed an unsupervised cosegmentation method in which the problem is formulated as an energy maximization problem that maximizes the similarity between candidate regions of the object in input images; they used random forest regression to obtain the similarity score. Meng et al. [5] also proposed an unsupervised cosegmentation method based on the shortest path problem. Hochbaum et al. [16] optimized an energy function in polynomial time using a maximum flow procedure. Rubio et al. [15] proposed a method based on Markov random fields to extract a common object from multiple images by

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806317.

Figure 2: Illustration of the process to transform an image into the graph. First column, second column and third column show original images, superpixels obtained by SLIC and the resultant graph.

minimizing an energy function that includes an inter-image region matching term.

Among the many strategies for the cosegmentation problem, graph matching has been known as one of the best performing approaches because of its high flexibility in matching deformable objects and regions. Yu et al. [14] proposed a graph-matching-based approach that transforms each input image into a graph and obtains foreground regions by a local graph matching algorithm that computes matches with regard to only one common region from the input graphs. This appears conceptually reasonable for the cosegmentation problem; however, a candidate foreground region obtained by local matching tends to include background slightly. Specifically, when visually similar backgrounds (e.g., sky, sea and grass) commonly appear across images, they are also matched and may be included in the estimated foreground.

In this paper, we propose an unsupervised cosegmentation algorithm based on global graph matching. Using a global graph matching algorithm that can compute both a small common subgraph and one-to-one mapping for every vertex between input graphs, we can find all paired regions and remove ones those that seem to be background to obtain a more accurate foreground region. Specifically, the subgraph of a foreground region is selected from candidate subgraphs using an algorithm based on connected component decomposition.

2. PROPOSED METHOD

The proposed method consists of the following four steps illustrated in Figure 1 (bottom):

(1) Transform each image into a graph.

(2) Find matches between graphs.

(3) Extract the subgraph representing the foreground.

(4) Refine the foreground region.

We describe the detail of each step in the following.

2.1 Transform each image into a graph

To suppress the number of vertices in a graph, we employ superpixel representations. For this purpose, we use the SLIC [6] algorithm introduced by Achanta et al. SLIC is based on K-means; however, it differs from typical K-means-based clustering in some aspects. For example, it moves centroids to the lowest gradient position in the initialization step and reduces the search region for clustering each superpixel.

Thus, all input images are transformed into superpixels, each of which is considered as a vertex of the graph. Each vertex is connected to its neighbors within two steps. The dot product between the mean vectors of SIFT features extracted from each superpixel is employed as the similarity between a pair of vertices.

Figure 2 shows superpixels obtained by SLIC (middle) and the resultant graph (right) of an image in the iCoseg [10] dataset.

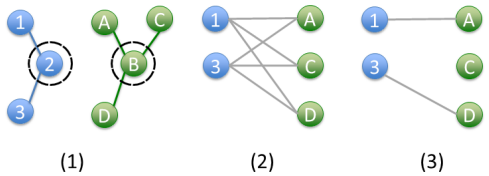

Figure 3: Coarse-grained phase process. (1) A vertex is chosen from each graph, and (2) the complete weighted bipartite graph is constructed from the neighbors of each vertex. (3) The maximum weighted matching score is considered as the similarity of these two vertices.

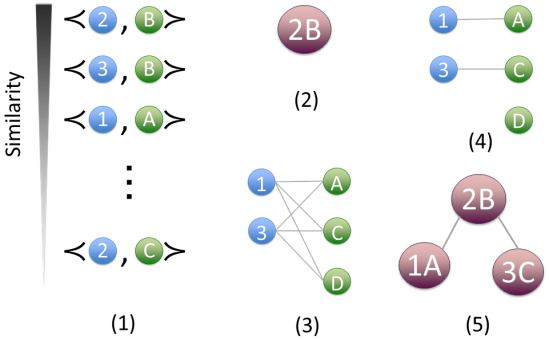

Figure 4: Fine-grained phase process. (1) Pairs of vertices are sorted in descending order of the similarity computed in the coarse-grained phase, and (2) the most similar pair is added to the alignment network. (3) The complete weighted bipartite graph is constructed from the neighbors of the vertices, and (4) the maximum weighted matching is computed. (5) The pairs of endpoints of the maximum weighted matching are added to the alignment network to increase the number of edges in E_{12}.

2.2 Find matches between graphs

We use the SPINAL algorithm [7] to find all matches from the graphs of a pair of images. SPINAL was originally proposed by Aladağ et al. to find matches between protein-protein interaction networks. In the field of bioinformatics, it has been suggested that SPINAL is more scalable than other algorithms such as IsoRank [8], MI-GRAAL [9], and NATALIE [17]. SPINAL is a global graph matching algorithm that computes one-to-one mapping for every vertex between two graphs.

Let $G_1 = (V_1, E_2)$ and $G_2 = (V_2, E_2)$ be two input graphs, where V and E represent their vertices and edges, respectively. To represent matches between graphs, alignment network $A_{12} = (V_{12}, E_{12})$ is defined. Each node of V_{12} represents a pair $\prec u_i, v_j \succ$, where $u_i \in V_1$ and $v_j \in V_2$. For

any pair of vertices $\prec u_i, v_j \succ \in V_{12}$ and $\prec u'_i, v'_j \succ \in V_{12}$ it should be the case that $u_i \neq u'_i$ and $v_j \neq v'_j$. The edge set of the alignment network is defined as $(\prec u_i, v_j \succ, \prec u'_i, v'_j \succ) \in E_{12}$ if $(u_i, u'_i) \in E_1$ and $(v_j, v'_j) \in E_2$. The goal of SPINAL is to find the alignment network A_{12} that maximize the global network alignment score (GNAS), which is defined as follows;

$$GNAS(A_{12}) = \alpha|E_{12}| + (1-\alpha) \sum_{\forall \prec u_i, v_j \succ} seq(u_i, v_j), \quad (1)$$

where $|E_{12}|$ denotes the number of edges in E_{12}, and $seq(u_i, v_j)$ denotes the similarity between two vertices, where $u_i \in V_1$ and $v_j \in V_2$. Here $\alpha \in [0, 1]$ is a parameter to balance the network-topological similarity and the sequence similarities. We set $\alpha = 0.5$ for all experiments in this study.

Generally, finding the optimal solution for the above problem is NP-hard [7]. To relax the problem efficiently, SPINAL consists of two main phases: a coarse-grained phase that computes rough similarities of all pairs of vertices $u_i \in V_1$ and $v_j \in V_2$, and a fine-grained phase that constructs the alignment network.

Figure 3 shows the process of the coarse-grained phase. We denote the set of neighbors of u_i in G_1 as $N(u_i)$ and the set of neighbors of v_j in G_2 as $N(v_j)$. In the coarse-grained phase, the neighborhood bipartite graph ($\mathcal{NBG}$), a complete edge-weighted bipartite graph defined on the partitions $N(u_i)$ and $N(v_j)$, is constructed and the similarity between $N(u_i)$ and $N(v_j)$ is computed by the maximum weighted matching of $\mathcal{NBG}$. In this manner, initial matching scores are computed for all pairs $u_i \in V_1$ and $v_j \in V_2$.

Figure 4 shows the process of the fine-grained phase. In this phase, the most similar pair of vertices not contained in V_{12} is first added to V_{12}. Then, its $\mathcal{NBG}$ is constructed, and the maximum weighted matching is added to V_{12}. For the endpoints of an edge not contained in the matching, their $\mathcal{NBG}$s are repeatedly constructed. This process is repeated until there is no pair of vertices not contained in V_{12}.

2.3 Extract the subgraph representing the foreground

In this step, we extract the subgraph that represents the common foreground from the alignment network obtained by SPINAL. A depth-first search is applied to the alignment network to decompose the connected components, and each connected component is evaluated by the objective function (1) of SPINAL. The connected component with the highest score is considered as the subgraph that represents the foreground.

Although it is not guaranteed that a subgraph extracted in this approach always corresponds to the foreground in general cases, our assumption is that the foreground occupies large areas in images, which we consider reasonable for many practical situations.

2.4 Refine the foreground region

To extract the foreground more precisely, we use GrabCut [12] proposed by Rother et al. Given small portions of foreground and background annotations as seeds, GrabCut estimates their entire regions. In our pipeline, we use the region corresponding to the subgraph obtained in the previous step as the foreground seed. In this manner, the coarse foreground obtained by graph matching is refined at the pixel level.

Table 1: Cosegmentation accuracy on iCoseg (%).

iCoseg	GrabCut	Rubio+[15]	Yu+[14]	Ours
Alaskan bear	84.3	86.4	78.0	**86.7**
Red sox players	85.4	90.5	86.4	**95.5**
Stonehenge 1	78.0	87.3	**87.4**	87.2
Stonehenge 2	70.7	**88.4**	74.7	87.7
Liverpool	75.4	82.6	**82.9**	82.4
Ferrari	79.7	84.3	86.1	**91.5**
Taj Mahal	82.5	**88.7**	85.4	79.0
Elephants	84.7	75.0	82.2	**95.5**
Pandas	68.3	60.0	76.8	**85.8**
Kite	86.4	89.8	85.2	**92.6**
Kite Panda	73.0	78.3	80.1	**86.7**
Gymnastics	85.0	87.1	90.7	**96.6**
Skating	65.5	76.8	77.3	**78.4**
Hot Balloons	83.2	89.0	86.3	**91.8**
Liberty Statue	73.1	**91.6**	87.3	88.2
Brown Bear	70.7	80.4	81.5	**95.6**
Average	77.9	83.9	83.0	**88.8**

Table 2: Cosegmentation accuracy on MSRC (%).

MSRC	Images	GrabCut	Rubio+[15]	Ours
Cars (front)	6	75.9	65.9	**84.2**
Cars (back)	6	72.5	52.4	**81.1**
Face	30	67.8	76.3	**79.6**
Cow	30	83.9	80.1	**94.3**
Cat	24	81.3	77.1	**89.7**
Plane	30	81.3	77.0	**87.4**
Bike	30	71.5	62.4	**72.1**
Average	-	76.3	70.2	**84.1**

3. EXPERIMENTAL RESULTS

To evaluate the proposed method, we compared it with two state-of-the-art unsupervised graph-matching-based algorithms [15, 14] [1] on the widely used iCoseg [10] (38 object classes; approximately 17 images per class) and MSRC [11] (14 classes; approximately 30 images per class) datasets, equipped with pixel-level segmentation ground truth. To ensure fair comparison with previous methods, we used a subset that consists of 16 classes from iCoseg and 7 classes from MSRC. Segmentation accuracy was evaluated in terms of the ratio of the number of correctly labeled pixels to the total number of pixels in an image. For each class, we randomly paired all images to use them as testing queries and reported the average accuracy. Tables 1 and 2 show experimental results for iCoseg and MSRC, respectively.

Table 1 shows the experimental results for iCoseg. The second column shows the results obtained using GrabCut only, and the third and fourth columns show the results obtained using previous methods. The last column shows the results obtained using the proposal method. Table 2 shows results for MSRC (the second column shows the number of images).

These results demonstrate that the proposed method outperforms previous methods in nearly all classes. As discussed in the introduction, previous methods find matches for only one visually common region between images that often undesirably include background. On the other hand, the proposed method finds matches for all vertices and removes the connected components that are dissimilar to be

[1]Some recent work has achieved very good performance by simultaneously cosegmenting more than two images [13, 2]. Since this is not included in our current scope, we focus on those using only two images as input.

Figure 5: Cosegmentation results on the iCoseg (top) and MSRC (bottom) datasets. The results of each class consist of four images. The two left images show the original images and the two right images show the common object.

foreground. Therefore, the proposed method achieves much significantly higher accuracy than the previous methods.

Compared with the state-of-the-art method [2] with cosegmentating of more than two images simultaneously, the results show higher accuracy than the proposed method (89.6% iCoseg). However, we expect to improve the accuracy of the proposed method by extending it to enable cosegmentation of multiple images.

Figure 5 shows some examples of images from the iCoseg and MSRC datasets, as well as their foreground regions extracted by the proposed method. These results illustrate that the proposed method can obtain proper foregrounds under difficult conditions such as miscellaneous backgrounds, viewpoint and scale changes, and object deformation.

4. CONCLUSIONS

We have proposed a method of unsupervised object cosegmentation based on a global graph matching algorithm. A local graph matching algorithm has been applied to find a common region in previous methods; however, the proposed method exploits global graph matching that can find a one-to-one mapping for every vertex between input graphs to remove background regions more accurately. Experimental results obtained with the iCoseg and MSRC datasets demonstrate the effectiveness of the proposed metehod. In addition, experimental results indicate that the accuracy of the proposed method is higher than that of previous methods.

In future, we will extend the proposed method to enable cosegmentation of more than two images or image streams.

5. REFERENCES

[1] X. Chen et al. NEIL: Extracting visual knowledge from Web data. In Proc. of IEEE ICCV, 2013.

[2] M. Rubinstein et al. Unsupervised joint object discovery and segmentation in internet images. In Proc. of IEEE CVPR, 2013.

[3] C. Rother et al. Cosegmentation of image pairs by histogram matching - Incorporating a global constraint into MRFs. In Proc. of IEEE CVPR, 2006.

[4] S. Vicente et al. Object cosegmentation. In Proc. of IEEE CVPR, 2011.

[5] F. Meng et al. Object co-segmentation based on shortest path algorithm and saliency model. ACM Trans. Multimedia, vol. 14, no. 5, pp. 1429-1441, 2012.

[6] R. Achanta, A. Shaji, K. Smith, A. Luchi, P. Fua, and S. Susstrunk. SLIC: Superpixels compared to state-of-the-art superpixel methods. IEEE Trans. PAMI, vol. 34, no. 11, pp. 2274-2281, 2012.

[7] A. E. Aladağ et al. SPINAL: scalable protein interaction network alignment. Bioinformatics, vol. 29, no. 7, pp. 917-924, 2013.

[8] R. Singh et al. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc. Natl. Acad. Sci. USA, vol. 105, no. 35, pp. 12763-12768, 2008.

[9] O. Kuchaiev et al. Integrative network alignment reveals large regions of global network similarity in yeast and human. Bioinformatics, vol. 27, no. 10, pp. 1390-1396, 2011.

[10] D. Batra et al. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In Proc. of IEEE CVPR, 2010.

[11] J. Winn et al. Object Categorization by Learned Universal Visual Dictionary. In Proc. of IEEE ICCV, 2005.

[12] C. Rother et al. "GrabCut" - Interactive foreground extraction using iterated graph cuts. In ACM Trans. Graphics, vol. 23, no. 3, pp. 309-314, 2004.

[13] A. Factor et al. Co-segmentation by composition. In Proc. of IEEE ICCV, 2013.

[14] H. Yu et al. Unsupervised cosegmentation based on superpixel matching and Fastgrabcut. In Proc. of IEEE ICME, 2014.

[15] J. C. Rubio et al. Unsupervised co-segmentation through region matching. In Proc. of IEEE CVPR, 2012.

[16] D. S. Hochbaum et al. An efficient algorithm for Co-segmentation. In Proc. of IEEE ICCV, 2009.

[17] G. W. Klau et al. A new graph-based method for pairwise global network alignment. BMC Bioinformatics, vol. 10, suppl. 1, S59, 2009.