GROUP SALIENCY PROPAGATION
FOR LARGE SCALE AND QUICK IMAGE CO-SEGMENTATION

Koteswar Rao Jerripothula*† *Jianfei Cai*† *Junsong Yuan*§

*ROSE Lab, Interdisciplinary Graduate School, Nanyang Technological University
†School of Computer Engineering, Nanyang Technological University
§School of Electrical and Electronic Engineering, Nanyang Technological University

ABSTRACT

Most of the existing co-segmentation methods are usually complex, and require pre-grouping of images, fine-tuning a few parameters and initial segmentation masks etc. These limitations become serious concerns for their application on large scale datasets. In this paper, Group Saliency Propagation (GSP) model is proposed where a single group saliency map is developed, which can be propagated to segment the entire group. In addition, it is also shown how a pool of these group saliency maps can help in quickly segmenting new input images. Experiments demonstrate that the proposed method can achieve competitive performance on several benchmark co-segmentation datasets including ImageNet, with the added advantage of speed up.

***Index Terms*—** co-segmentation, propagation, quick, large scale, group, ImageNet, fusion.

1. INTRODUCTION

Image co-segmentation deals with segmenting out common objects from a set of images. Exploiting the weak supervision provided by association of similar images, co-segmentation has become a promising substitute to tedious interactive segmentation for object extraction and labelling in large-scale datasets. The concept was originally introduced by [1] to simultaneously segment out the common object from a pair of images. Since then, many co-segmentation methods have been proposed to scale from image pair to multiple images [2, 3, 4]. The work in [2] proposed a discriminative clustering framework and [3] used optimization for co-segmentation. Recently, [5] combined co-segmentation with co-sketch for effective co-segmentation and [6] proposed to co-segment a noisy image dataset (where some images may not contain the common object). Most of these existing works require intensive computation due to co-labelling multiple images and also need fine-tuning a few parameters and pre-grouping of images etc.

Our previous work [7] addresses some of these issues by performing segmentation on each of the individual images but with the help of a global saliency map (formed *via* pairwise warping process). Warping [8] is a process of alignment of one image w.r.t. another image where dense correspondence between two images is established at pixel level and this process is computationally expensive. Therefore, developing such a global saliency map for each image requires intensive computation. In this paper, we aim at improving the efficiency of our previous method so that it can be applied to large scale datasets without much performance loss.

Geometric Mean Saliency: Since proposed method is based on our previous work [7], we briefly review the basic idea here to make the paper self contained. In GMS [7], segmentation on individual images is performed, but using a global saliency map that is obtained by fusing single-image saliency maps of a group of similar images. In this way, even if an existing single-image saliency detection method fails to detect the common object as salient in an image, saliency maps of other images can help in extracting the common object by forming a global saliency map. The fusion takes place at the pixel level using geometric mean after saliency maps of other images are aligned. The corresponding saliency values in other images are collected and combined for each image. When we do this for each image, mostly the same corresponding pixels get together every time and their saliency values are combined, under the assumption that warping technique is precise and accurate. This makes the process repetitive and inefficient. So, we avoid such repetition in the proposed method by doing this task only once and propagate this information in the group. Therefore, in this paper, we effectively perform the same task as earlier in an efficient way and as a result we hardly observe any drop in the performance.

In particular, our basic idea is to first cluster visually similar images into a group and select a key image to represent each group. After that we can align all the saliency maps of other images in the group w.r.t. the key image *via* warping technique and fuse them to form a single group saliency

$\star\dagger$This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Prime Ministers Office, Singapore, under its IDM Futures Funding Initiative and administered by the Interactive and Digital Media Programme Office. This research is also supported by Singapore MoE AcRF Tier-1 Grant RG30/11.

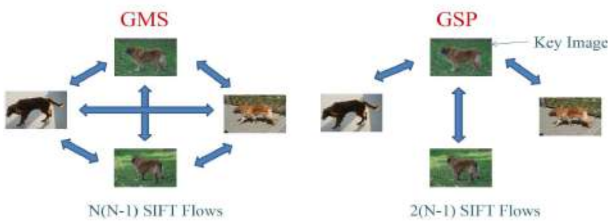

Fig. 1. Comparison between GMS [7] and our proposed GSP. Instead of performing pairwise matching (GMS [7]), GSP only matches each member image with the key image that represents the whole group.

map for the entire group. This group saliency map serves as a group prior for foreground extraction of all the images in the group by means of propagation that is done *via* warping. Moreover, it can also help to segment a new input image by the same means. We call our method Group Saliency Propagation (GSP) model. Note that there are only a few attempts on large-scale cosegmentation such as [3, 9, 10], where [10] represents state-of-the-art. Although [10] also relies on propagation, it needs some *human annotated segmentation masks and semantic grouping* to initiate the propagation, whereas our method does not have such a limitation. In addition, our GSP method can also help on quick segmentation of any new image by utilizing a pool of pre-computed group saliency maps. Experiments on several benchmark datasets including ImageNet [11], show that our proposed GSP model can achieve competitive results with a drastic speed up.

2. GROUP SALIENCY PROPAGATION

This section provides the detailed description of our proposed method. The group saliency map is obtained by fusing the key image saliency map and warped saliency maps of other images at pixel level. Unlike [7] where such fusion of individual saliency map and warped saliency map is carried out for every image, here we only fuse the saliency maps for the key image and treat the fused saliency map as the group saliency map. Once we obtain the group saliency map for each group, this group saliency map can help in segmentation of all the images in the group as well as new input images through propagation.

Notations: Consider a large dataset with m images $\mathbf{I} = \{I_1, I_2, .., I_m\}$. Let $\mathbf{M} = \{M_{I_1}, M_{I_2}, .., M_{I_m}\}$ be the set of corresponding visual saliency maps, $\mathbf{D} = \{D^{I_1}, D^{I_2}, ..., D^{I_m}\}$ be the set of corresponding weighted GIST descriptors [12, 6] (saliency maps are used as weight maps) for global description, and $\mathbf{L} = \{L^{I_1}, L^{I_2}, ..., L^{I_m}\}$ be the set of corresponding dense SIFT descriptors [8] for local description. Denote $\mathcal{I}, \mathcal{A}, \mathcal{S}$ and $\mathcal{M}$ as a new input image, its GIST descriptor, SIFT descriptor and preprocessed saliency maps respectively.

Pre-processing: We use the same pre-processing steps as in [7] to ensure that each saliency map covers sufficient regions of the object by adding spatial contrast saliency and then brighten the saliency map to avoid over-penalty caused by low saliency values.

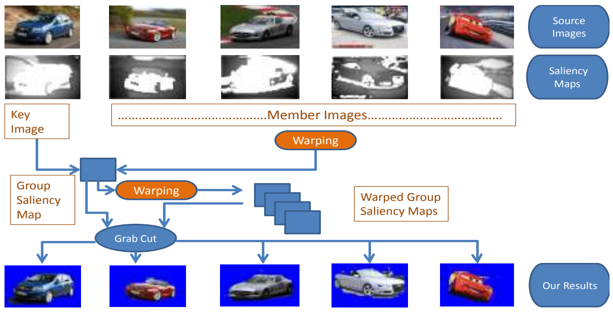
Source Images
Saliency Maps
Key Image
....Member Images..
Warping
Group Saliency Map
Warping
Warped Group Saliency Maps
Grab Cut
Our Results

Fig. 2. Proposed GSP approach: A single group saliency map of key image helps in segmenting all the other images.

2.1. Group Saliency Maps

We first perform k-means clustering using GIST descriptors of images to cluster the entire image dataset into K groups. For each group the image closest to a cluster center is selected as the key image to represent the whole group. For each group, its group saliency map is produced using [7] w.r.t the key image alone where saliency maps of the member images are aligned to the key image *via* the warping technique. Then following the method in GMS [7], the key image saliency map and the warped saliency maps from other members are fused at pixel level using geometric mean. Let C_k denote the k-th cluster as well as the corresponding key image, and $U^{I_j}_{C_k}$ denote the warped saliency map of the image I_j w.r.t. key image C_k. The group saliency map G^k value for any pixel p can now be computed as

$$G^k(p) = \left(M_{C_k}(p) \prod_{I_j \neq C_k, I_j \in C_k} U^{I_j}_{C_k}(p) \right)^{\frac{1}{|C_k|}} \qquad (1)$$

where $|.|$ represents cardinality.

For quick co-segmentation, along with the group saliency maps we also need to store GIST and SIFT descriptors of the key images and number of members in the group for the purpose of group assignment, warping and fusion respectively. Therefore, the pool of group saliency maps contains not only the maps $\mathbf{G} = \{G^1, G^2, ..., G^K\}$ but also the corresponding GIST and SIFT descriptors and number of group members: $\mathbf{A} = \{D^{C_1}, D^{C_2}, ..., D^{C_K}\}$ and $\mathbf{S} = \{L^{C_1}, L^{C_2}, ..., L^{C_K}\}$ and $\mathbf{C} = \{|C_k|, |C_2|, ..., |C_K|\}$ respectively.

2.2. Propagation

In our previous work [7], for a group of n images, when segmenting one image, the saliency maps of the other $n-1$ images need to wrap to the current image, which requires computing the costly dense SIFT flows $n-1$ times. Thus, seg-

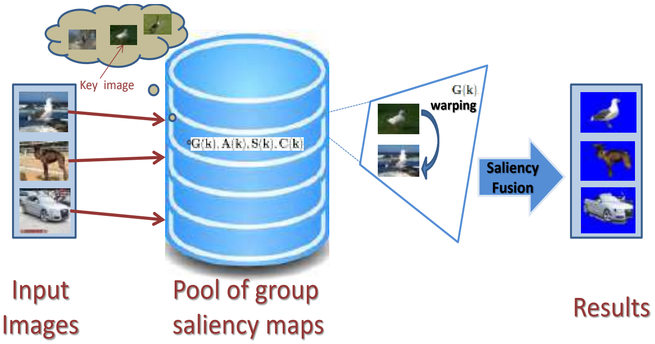

Fig. 3. Quick co-segmentation of any new input images by first group assigment, then group saliency map warping and then fusion of saliency maps.

menting n images will need computing dense SIFT flows for $n \times (n-1)$ times, as shown in the left part of Fig. 1, which is very time consuming, especially for large-scale dataset.

In order to reduce the computational cost, in this paper, we propose to use group saliency map to propagate the group prior information. In particular, for a cluster of n images, we only compute the dense SIFT flows for the key image $n-1$ times. When segmenting the $n-1$ member images, we simply warp the group saliency map back to each of them. In this way, totally we only need $2 \times (n-1)$ dense SIFT flows, as shown in the right part of Fig. 1. We achieve a drastic speed up of $n/2$ times by using this kind of propagation of group saliency maps. An illustration of the proposed approach is given in Fig. 2. In particular, the object prior $\mathcal{O}$ for any member image I_j would be warped G^k w.r.t. I_j denoted as $U_{I_j}^{C_k}$.

Based on the pool of $[\mathbf{G}, \mathbf{A}, \mathbf{S}, \mathbf{C}]$ information, we can also quickly co-segment a new image by 1) assigning the image to a group using the GIST feature, 2) warping the selected group saliency map to the input image, and 3) fusing the saliency map of the input image with the warped group saliency map, as shown in Fig. 3. Since one more saliency value is to be included in the calculation of geometric mean of saliency values, the object prior $\mathcal{O}$ value for any pixel p of a new input image $\mathcal{I}$ can be computed as,

$$\mathcal{O}(p) = \left(\left(U_{\mathcal{I}}^{C_k}(p) \right)^{|C_k|} \times \mathcal{M}(p) \right)^{\frac{1}{|C_k|+1}} \tag{2}$$

where $U_{\mathcal{I}}^{C_k}$ denotes warped G^k w.r.t. $\mathcal{I}$

2.3. Foreground Extraction

We obtain the final mask using GrabCut [13], in which regularization is performed with the help of SLIC superpixel oversegmentation [14]. Particularly, foreground (F) and background (B) seed locations for GrabCut are determined by

$$p \in \begin{cases} F, & \text{if } \mathcal{O}(p) > \tau \\ B, & \text{if } \mathcal{O}(p) < \phi \end{cases} \tag{3}$$

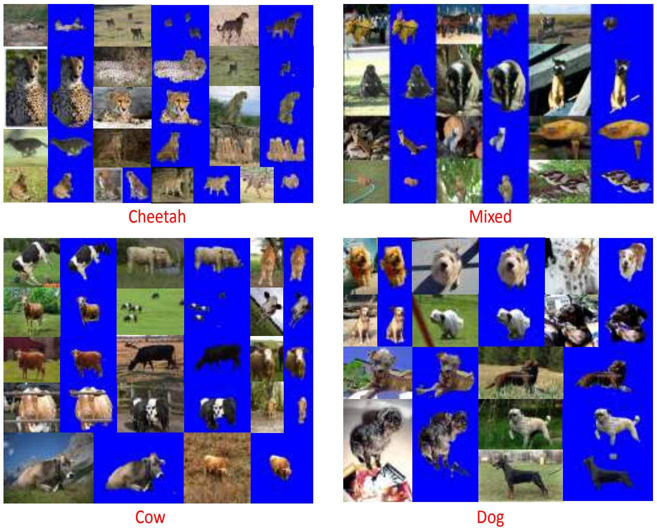

Fig. 4. Sample Results on ImageNet [11]

Table 1. Jaccard Similarity and Time consumed comparison of GSP with GMS[7] using default setting

		Jaccard Similarity		Time(mins.) Consumed	
Datasets	Images	GMS[7]	GSP	GMS[7]	GSP
Weizmann Horses	328	0.72	0.72	117	45
MSRC	418	0.68	0.68	105	22
iCoseg	529	0.67	0.66	126	38
CosegRep	572	0.71	0.71	216	35
Internet Images	2470	0.62	0.61	806	133

where ϕ is a global threshold value of $\mathcal{O}$ which is automatically determined by the common Otsu's method [15] and τ is the only tuning parameter.

3. EXPERIMENTAL RESULTS

Several benchmark datasets including ImageNet [11] have been used to validate the proposed GSP model and compare it with the existing methods. Two objective measures, Jaccard Similarity (J) and Precision (P), are used for evaluation. Jaccard Similarity is defined as the intersection divided by the union of ground truth and segmentation result, and Precision is defined as the percentage of pixels correctly labeled.

In Table 1, we compare our quantitative results and time consumed with our previous work [7] on 5 benchmark cosegmentation datasets: MSRC [17], iCoseg [18], Coseg-Rep [5], Internet Dataset [6], Weizmann Horses [19], using default parameter setting :$\tau = 0.99$. In each of our experiments with m images, K is calculated as nearest integer to $m/15$ en-

Table 2. Quantitative Comparison with existing methods on large scale dataset ImageNet

Methods	Jaccard Similarity	Precision
[10]	-	77.3
[16]	0.57	84.3
Ours(1)	0.55	83.9
Ours(2)	0.58	85.6
Ours(3)	0.62	87.7

Table 3. Comparison of using categorization and not using categorization in default setting scenario

	Yes		No	
Datasets	J	P	J	P
MSRC	0.678	87.2	0.674	86.9
iCoseg	0.656	88.9	0.632	87.8
CosegRep	0.706	91.2	0.701	90.4

suring that there are 15 images per sub-group on an average. Since apart from Weizmann horses dataset, all other datasets contain multiple sub-groups, experiments are conducted separately on each sub-group and average performance has been reported. Reported “Time Consumed” is the total time consumed for entire dataset. It can be noted that performance is hardly affected but there is significant improvement in speed. As explained in previous sections, the reason can be attributed to the fact that we are effectively achieving the same objective ,i.e., collecting corresponding pixels together and combining their saliency values. However, here we efficiently avoid the repetition occurring in our previous work [7].

Recently, evaluation on large dataset such as ImageNet has also become possible with the release of groundtruths on nearly 4.5k images by [10]. We apply our method on about 0.4 million images of 446 synsets to which these 4.5k images belong and compare our performance with [10] and [16] in Table 2. Here τ is the only parameter that we tune in our experiments and we report three results with different setting of τ: (1) τ is fixed for all classes, (2) τ is fixed within a class, but varied across the classes and (3) τ is tuned across all the images of dataset. As expected, the performance improves from (1) to (3) progressively and it will be interesting to see if this parameter can be learned. Compared with [16] our method does not require any segmentation masks for initiation but still achieves comparable performance as in [16]. Some sample segmentation results obtained from ImageNet are shown in Fig. 4. Even with cluttered backgrounds, our method can still segment the foregrounds successfully.

Since we rely on good clustering to group visually similar images and claim that manually pre-grouping is not absolutely essential for our method, it would be interesting to see how our method behaves “without exploiting manual categorization” in comparison to “with exploiting manual categorization” on benchmark co-segmentation datasets. It can be noted in Table 3 that there is very minor drop in performance.

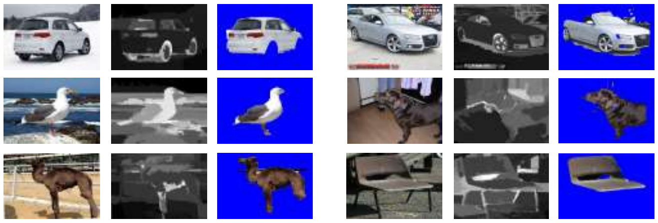

Fig. 5. Some sample results of our quick co-segmentation method. In spite of object not being so salient in the saliency map [20], the proposed approach is able to extract it.

Table 4. Jaccard similarity comparison beween quick co-segmentation and other automatic methods like GrabCut initialized by central window and saliency map [20]

Datasets	Center	Saliency	Quick
MSRC	0.44	0.47	**0.67**
Weizmann Horses	0.53	0.58	**0.72**

To evauate our quick cosegmentation idea we first make a pool of 30k group saliency maps produced while attempting to segment ImageNet and then input the images of datasets image-by-image. In Table 4, we compare results of our quick co-segmentation idea with that of simple GrabCut initialized with central window of 25% of size of the image and initialized with saliency prior thresholded using Otsu method. Grabcut initialized by central window, grabcut initialized by saliency prior and our quick co-segmentation method have been labelled as Center, Saliency and Quick respectively. It can be noted that our quick co-segmentation method obtains significant improvement over others. Some sample results of this approach are shown in Fig. 5 demonstrating how foreground gets extracted in spite of not being so salient.

4. CONCLUSION

Image co-segmentation in large-scale dataset remaining a challenging problem as most existing methods cannot be easily extended to big dataset, we efficiently extend our previous work [7] to perform large-scale co-segmentation and quick co-segmentation of new input images by introducing the idea of propagating a group saliency map in the entire group. Such an approach requires a lesser number of SIFT flows compared to the previous method. Experimental results demonstrated that proposed method is able to perform co-segmentation in a much faster manner without much affecting on performance.

5. REFERENCES

[1] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov, “Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs,” in *Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2006, vol. 1, pp. 993–1000.

[2] Armand Joulin, Francis Bach, and Jean Ponce, “Discriminative clustering for image co-segmentation,” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 1943–1950.

[3] Gunhee Kim, Eric P Xing, Li Fei-Fei, and Takeo Kanade, “Distributed cosegmentation via submodular optimization on anisotropic diffusion,” in *International Conference on Computer Vision(ICCV)*. IEEE, 2011, pp. 169–176.

[4] Armand Joulin, Francis Bach, and Jean Ponce, “Multi-class cosegmentation,” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 542–549.

[5] Jifeng Dai, Ying N Wu, Jie Zhou, and Song-Chun Zhu, “Cosegmentation and cosketch by unsupervised learning,” in *International Conference on Computer Vision(ICCV)*, 2013.

[6] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu, “Unsupervised joint object discovery and segmentation in internet images,” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 1939–1946.

[7] Koteswar R Jerripothula, Jianfei Cai, Fanman Meng, and Junsong Yuan, “Automatic image Co-Segmentation using geometric mean saliency,” in *International Conference on Image Processing (ICIP)*, Paris, France, Oct. 2014, pp. 3282–3286.

[8] Ce Liu, Jenny Yuen, and Antonio Torralba, “Sift flow: Dense correspondence across scenes and its applications,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 5, pp. 978–994, 2011.

[9] Gunhee Kim and Eric P Xing, “On multiple foreground cosegmentation,” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 837–844.

[10] Daniel Kuettel, Matthieu Guillaumin, and Vittorio Ferrari, “Segmentation propagation in imagenet,” in *European Conference on Computer Vision(ECCV)*, pp. 459–473. Springer, 2012.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[12] Aude Oliva and Antonio Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International journal of computer vision(IJCV)*, vol. 42, no. 3, pp. 145–175, 2001.

[13] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, “Grabcut: Interactive foreground extraction using iterated graph cuts,” in *Transactions on Graphics (TOG)*. ACM, 2004, vol. 23, pp. 309–314.

[14] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, “Slic superpixels,” *Ecole Polytechnique Fédéral de Lausssanne (EPFL), Tech. Rep*, vol. 2, pp. 3, 2010.

[15] Nobuyuki Otsu, “A threshold selection method from gray-level histograms,” *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[16] Matthieu Guillaumin, Daniel Kuettel, and Vittorio Ferrari, “Imagenet auto-annotation with segmentation propagation,” *International Journal of Computer Vision(IJCV)*, pp. 1–21, 2014.

[17] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi, “Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation,” in *European Conference on Computer Vision(ECCV)*, pp. 1–15. Springer, 2006.

[18] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen, “icoseg: Interactive co-segmentation with intelligent scribble guidance,” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3169–3176.

[19] Eran Borenstein and Shimon Ullman, “Class-specific, top-down segmentation,” in *European Conference on Computer Vision(ECCV)*, pp. 109–122. Springer, 2002.

[20] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu, “Global contrast based salient region detection,” in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 409–416.