AUTOMATIC IMAGE CO-SEGMENTATION USING GEOMETRIC MEAN SALIENCY

Koteswar Rao Jerripothula[⋆†] *Jianfei Cai*[†] *Fanman Meng*[†] *Junsong Yuan*[§]

[⋆] ROSE Lab, Interdisciplinary Graduate School, Nanyang Technological University, Singapore
[†] School of Computer Engineering, Nanyang Technological University, Singapore
[§] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

Most existing high-performance co-segmentation algorithms are usually complicated due to the way of co-labelling a set of images and the requirement to handle quite a few parameters for effective co-segmentation. In this paper, instead of relying on the complex process of co-labelling multiple images, we perform segmentation on individual images but based on a combined saliency map that is obtained by fusing single-image saliency maps of a group of similar images. Particularly, a new multiple image based saliency map extraction, namely geometric mean saliency (GMS) method, is proposed to obtain the global saliency maps. In GMS, we transmit the saliency information among the images using the warping technique. Experiments show that our method is able to outperform state-of-the-art methods on three benchmark co-segmentation datasets.

Index Terms— co-segmentation, image segmentation, saliency, warping.

1. INTRODUCTION

Image co-segmentation has drawn a lot of attention from vision community as it can provide unsupervised information regarding "what to segment out" with the help of other images that contain similar object. The concept was first introduced by Rother et al. [1], who used histogram matching to simultaneously segment out the common object from a pair of images. Since then, many co-segmentation methods have been proposed to either improve the segmentation in terms of accuracy and processing speed [2, 3, 4, 5, 6, 7] or scale from image pair to multiple images [4, 8, 9]. Joulin et al. [4] proposed a discriminative clustering framework and Kim et al. [8] used optimization for co-segmentation. Dai et al. [10] combined co-segmentation with cosketch for effective co-segmentation. Recently, Rubinstein et al. [11] adopted dense SIFT matching to discover common objects and co-segment them out from noisy image dataset (where some images do not contain the common object).

⋆This research is supported by the Singapore National Research Foundation under its IDM Futures Funding Initiative and administered by the Interactive & Digital Media Programme Office, Media Development Authority.

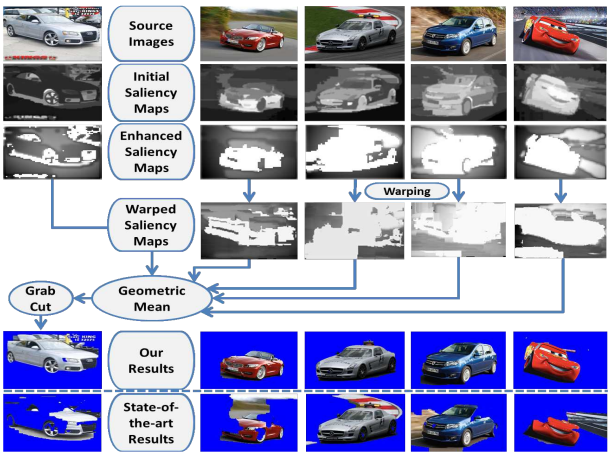

Fig. 1. Proposed GMS method, where salient common object images can render help to weakly salient common object image (the first image) in segmentation. The bottom row shows the results of state-of-the-art co-segmentation method [11].

Despite the previous progress, there still exist some major problems for the existing co-segmentation algorithms. 1) As shown in [12, 11], co-segmenting images might not perform better than single-image segmentation for some datasets. This raises up the question: to co-segment or not. 2) Most of the existing high-performance co-segmentation algorithms are usually complicated due to the way of co-labelling a set of images, and require handling quite a few parameters for effective co-segmentation, which becomes more difficult when the dataset becomes increasingly diverse.

In this paper, instead of relying on the complex process of co-labelling multiple images, we perform segmentation on individual images but using a combined 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. Fig. 1 demonstrates how the first image containing weakly salient common object (car) is helped by images containing similar salient common object (car).

In particular, we first enhance individual saliency maps to highlight the foreground. We then group the original images into several clusters using GIST descriptor [13]. After that, Dense SIFT flow [14] is computed to obtain pixel correspondence between image pairs within each cluster, which is used to obtain the warped saliency maps. Enhanced saliency maps and warped saliency maps are combined to obtain the global saliency maps. Based on the global saliency maps, we select the foreground and background seeds and use GrabCut [15] to segment out the common object. The proposed method is verified on three public datasets (MSRC[16], iCoseg [2], Coseg-Rep [10]). The experimental results show that the proposed method can obtain larger IOU (Intersection over Union) values compared with state-of-the-art methods.

2. PROPOSED METHOD

The basic idea of our method is to cluster similar images into subgroups, use dense correspondence to transfer saliency among similar images in a subgroup, and finally perform individual image segmentation based on the combined saliency map. Fig.2 shows the flow chart of our proposed method. The detailed procedure is explained in the following subsections.

2.1. Notation

Let $I = \{I_1, I_2, \cdots, I_m\}$ be the image group containing m images, D_i be the image domain of I_i, $p \in D_i$ be pixel with coordinates (x, y). The geometric mean saliency maps are denoted as $G = \{G_1, G_2, \cdots, G_m\}$. The foreground and background seeds are represented as F_i and B_i for image I_i, respectively.

2.2. Saliency Enhancement

We use the method in [17] to obtain the initial saliency maps, which are denoted as $L = \{L_1, L_2, \cdots, L_m\}$. In this step, we first obtain the binary map T_i for L_i by a global threshold t_i (computed by the conventional Otsu's method[18]) , i.e., $T_i(p) = 1$ when $L_i(p) > t_i$; otherwise, $T_i(p) = 0$. To ensure that a saliency map covers sufficient regions of the salient object, we further enhance the saliency values of some background pixels. Particularly, for the background pixel p (with value zero in T_i), we compute its spatial contrast saliency value $T_i'(p)$ by

$$T_i'(p) = \delta(T_i(p) = 0) \sum_{q \in D_i} |I_i'(p) - I_i'(q)| e^{\frac{-d_{pq}}{\sigma}} \quad (1)$$

where $\delta(\cdot) = 1$ only if $\cdot$ is true, I_i' is the gray image of I_i and d_{pq} is the distance between the location of pixels p and q. σ is set to 25. T_i' is then normalized and set as the new value for the background pixels, i.e., $T_i(p) = T_i'(p)$ when $T_i(p) = 0$, making T_i a continuous map. To speed up the computation, this step is performed on downsized images (say 50x50).

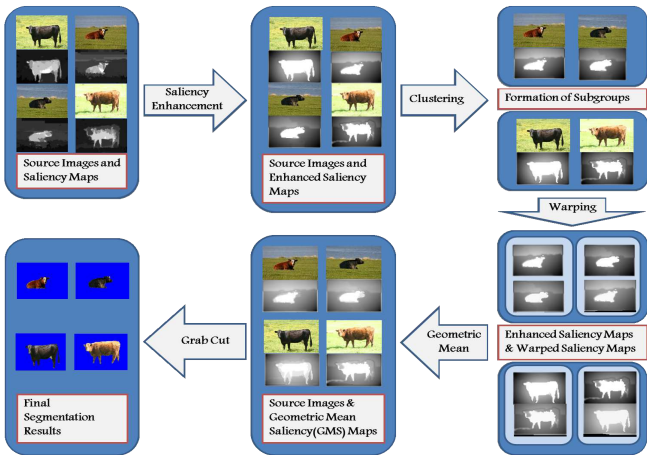

Fig. 2. Flow chart of proposed method

To facilitate the later geometric mean operation, brighter saliency maps are preferred so as to avoid over-penalty caused by low saliency values. Thus, we perform log transformation to make the saliency map brighter, i.e.,

$$M_i(p) = \log_{(1+\mu)} (1 + \mu T_i(p)) \quad (2)$$

where μ is set to 300 and M_i is the enhanced saliency map.

2.3. Subgroup Formation

After enhancing the saliency map, we next cluster the images into a number of image subgroups, where the images within the same subgroup have similar appearance. Here, weighted GIST descriptor [13, 11] is used to represent each image with enhanced saliency map M_i as the weights for image I_i. The k-means clustering algorithm is applied for clustering. Let K be the number of clusters and C_k be the set of indexes of images in k^{th} cluster, where $k \in \{1, \cdots, K\}$. In general, 10 images per group are good enough for our model, K is determined according to the total number of images m, i.e., K is calculated as nearest integer to $m/10$.

2.4. Pixel Correspondence

Based on the clustering results, we intend to match the pixels among the images within each subgroup. Specifically the masked Dense SIFT correspondence [14, 11] is used to find corresponding pixels in each image pair. We obtain the masks by thresholding saliency maps M's. Let the masks from M_i and M_j be T_{M_i} and T_{M_j} respectively. The objective function for Dense SIFT flow is represented as

$$E(w_{ij}; T_{M_i}, T_{M_j}) = \sum_{p \in D_i} T_{M_i}(p) \Big(T_{M_j}(p + w_{ij}(p)) ||S_i(p) - S_j(p + w_{ij}(p))||_1 + \big(1 - T_{M_j}(p + w_{ij}(p)\big) C_0 + \sum_{q \in N_p^i} \alpha ||w_{ij}(p) - w_{ij}(q)||_2 \Big) \quad (3)$$

Fig. 3. Sample segmentation results for iCoseg dataset .

Fig. 4. Sample segmentation results for Repititive category of Coseg-Rep dataset

where S_i is dense SIFT descriptor for image I_i, N_p^i is neighborhood of p, α weighs smoothness, C_0 is a large constant, and w_{ij} denotes flow field from image I_i to I_j.

2.5. Geometric Mean Saliency

Given a pair of images I_i and I_j from a subgroup C_k, we transmit the saliency map M_j to I_i through warping technique. We form the warped saliency map U_i^j by $U_i^j(p) = M_j(p')$, where (p, p') is a matched pair in the SIFT flow alignment with relationship $p' = p + w_{ij}(p)$. Since there are a few images in subgroup C_k, for each image we fuse their warped saliency maps along with its own saliency map by computing the GMS score $G_i(p)$,

$$G_i(p) = \sqrt[|C_k|]{M_i(p) \prod_{j \neq i}^{j \in C_k} U_i^j(p)} \tag{4}$$

where $|C_k|$ denotes number of images in C_k subgroup and GMS score is essentially the geometric mean of all the involved saliency maps.

2.6. Image Segmentation

Based on the GMS scores, we obtain the final mask using GrabCut algorithm [15], in which foregrounds and background seed locations are determined by

$$p \in \begin{cases} F_i, & \text{if } G_i(p) > \tau \\ B_i, & \text{if } G_i(p) < \phi_i \end{cases} \tag{5}$$

where ϕ_i is a global threshold value of G_i determined by the common Otsu's method[18] and τ is a parameter. Note that we also use regularization to make the GMS score consistent within a region. Specifically, the SLIC algorithm [19] is adopted to generate superpixels, and then each pixel's GMS score is replaced by the average GMS score of its corresponding superpixel.

Table 1. Comparison with state-of-the-art methods on MSRC and iCoseg datasets using overall values of J and P

	MSRC		iCoseg	
	J	P	J	P
Distributed[8]	0.37	54.7	0.40	70.4
Discriminative [4]	0.45	70.8	0.39	61.0
Multi-Class [9]	0.51	73.6	0.43	70.2
Object Discovery [11]	0.68	87.7	0.69	89.8
Proposed Method	**0.70**	**88.4**	**0.72**	**91.6**

3. EXPERIMENTAL RESULTS

Several challenging datasets including MSRC[16], iCoseg[2] and Coseg-Rep [10] are used in our experiments. Two objective measures, Precision (P) and Jaccard Similarity (J or IOU), are used for the evaluation. Precision is defined as the percentage of pixels correctly labelled, and Jaccard Similarity is defined as the intersection divided by the union of ground truth and segmentation result.

We only tune the parameter τ in the range [0.94,0.99] for each category in the datasets, and for other parameters, we use a fixed global setting: $\mu = 300$ and $\sigma = 25$. We compare with the methods [11, 8, 4, 9] on iCoseg and MSRC dataset and the method [10] on Coseg-Rep dataset.

The quantitative results (average J and P over all the categories) and the classwise comparisons results are displayed in Tables 1-2 and Fig. 6 respectively. The segmentation results of methods [8, 4, 9, 11] are taken from the experimental setup of Object Discovery[11]. It can be seen that the proposed method obtains larger J and P values than the state-of-the-art method [11] on MSRC and iCoseg dataset and state-of-the-art method [10] on Coseg-Rep dataset. Sample results for iCoseg and Coseg-Rep dataset are shown in Fig. 3 and Fig. 4 respectively. Our method outperforms [11] on 11/14 categories in MSRC dataset and 20/30 categories in iCoseg

(a)Image (b)[9] (c)[11] (d)Ours

Fig. 5. Comparison with state-of-the-art methods on cat and dog categories of MSRC dataset.

Table 2. Comparison with [10] on Coseg-Rep dataset

	J	P
Cosegmentaton&Cosketch[10]	0.67	90.2
Proposed Method	**0.73**	**92.2**

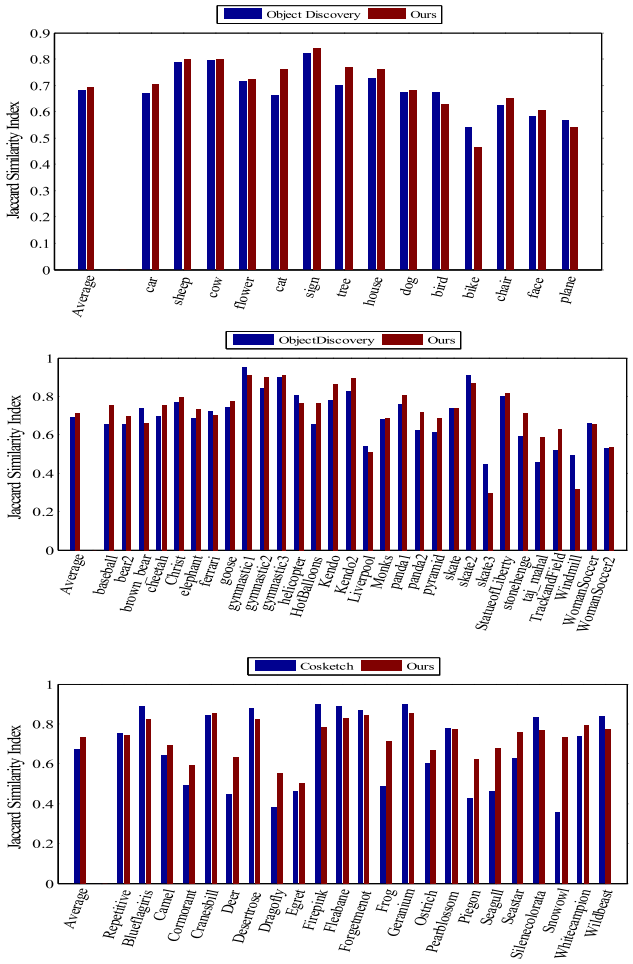

Fig. 6. Classwise Comparison with Object Discovery method [11] on MSRC and iCoseg datasets and with Cosketch method [10] on CosegRep datset

Dataset. Significant improvement is obtained on the Coseg-Rep dataset, where our method outperforms method [10] in 13/23 categories. For practical applications, our model performs sufficiently good with default parameter setting itself ($\mu = 300, \sigma = 25, \tau = 0.97$), for which we obtained **0.68**, **0.67** and **0.71** Jaccard Similarity on MSRC, iCoseg, Coseg-Rep datasets respectively. Fig. 5 gives visual comparision of the results of different methods for the MSRC dataset.

Another experiment we conduct is to merge all the categories of MSRC dataset into one category and verify if our model can be used on a mixed dataset. The experimental results show that the proposed method can obtain J as **0.68** again with default parameters itself. Note that although [11] also reports a J value of 0.68 on MSRC dataset, it tunes its parameters and performs co-segmentation for individual categories. This experiment demonstrates that our method can effectively handle mixed dataset with great diversity and achieve segmentation results that are close to that of state-of-the-art method [11] obtained by doing class-by-class co-segmentation.

4. CONCLUSION

We have proposed a saliency based automatic image co-segmentation method. Our main idea is to form a global saliency map for each image by fusing individual single-image saliency maps and then use the global saliency map to perform single-image segmentation. The experimental results demonstrate that by adjusting only one parameter per category our method can achieve the best performance in all the benchmark datasets, default parameters setting in our model produces sufficiently good results and proposed method can handle mixed dataset efficiently. Future works include extending the model to perform co-segmentation on noisy image dataset and large-scale dataset.

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, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 993–1000.

[2] 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), 2010 IEEE Conference on*. IEEE, 2010, pp. 3169–3176.

[3] Dorit S Hochbaum and Vikas Singh, "An efficient algorithm for co-segmentation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 269–276.

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

[5] Lopamudra Mukherjee, Vikas Singh, and Chuck R Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2028–2035.

[6] Junsong Yuan, Gangqiang Zhao, Yun Fu, Zhu Li, Aggelos K Katsaggelos, and Ying Wu, "Discovering thematic objects in image collections and videos," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 2207–2219, 2012.

[7] Gangqiang Zhao and Junsong Yuan, "Mining and cropping common objects from images," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 975–978.

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

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

[10] Jifeng Dai, Ying Nian Wu, Jie Zhou, and Song-Chun Zhu, "Cosegmentation and cosketch by unsupervised learning," in *14th International Conference on Computer Vision*, 2013.

[11] 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), 2013 IEEE Conference on*. IEEE, 2013, pp. 1939–1946.

[12] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov, "Object cosegmentation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2217–2224.

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

[14] 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.

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

[16] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Computer Vision–ECCV 2006*, pp. 1–15. Springer, 2006.

[17] 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), 2011 IEEE Conference on*. IEEE, 2011, pp. 409–416.

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

[19] 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.