21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Scalable Image Co-Segmentation Using Color and Covariance Features

Shijie Zhang[1], Wei Feng[1]*, Liang Wan[2], Jiawan Zhang[2], Jianmin Jiang[1]
[1]*School of Computer Science and Technology, Tianjin University, Tianjin, China*
[2]*School of Computer Software, Tianjin University, Tianjin, China*
{shijiezhang,wfeng,lwan,jwzhang,jmjiang}@tju.edu.cn

Abstract

This paper focuses on producing fast and accurate co-segmentation to a pair of images that is scalable and able to apply multimodal features. We present a general solution for this purpose and specifically propose a non-iterative and fully unsupervised method using pointwise color and regional covariance features for image co-segmentation. The scalability and generality of our method mainly attribute to the superpixel-level irregular graph formulation and multi-feature joint clustering. Through a unified similarity metric, the contributions of multiple features are finally embodied into the co-segmentation energy function. Experiments on common dataset validate the superior scalability of our method over state-of-the-art alternatives and its capability of generating comparable or even better labeling accuracy at the same time. We also find that multi-feature co-segmentation usually produces better labeling accuracy than using single color feature only.

1. Introduction

Aiming at jointly segmenting the common foreground regions from image pairs [9], co-segmentation is very useful in many semantic labeling tasks [3]. In the meantime, it is also theoretically important by providing an unsupervised manner to reduce the ambiguity of automatic foreground/background separation [11].

Co-segmentation is usually formalized as an energy minimization problem with the following energy form:

$$\mathrm{E}_{\mathrm{coseg}}(X) = \sum_{i=1}^{2} \mathrm{E}_{\mathrm{seg}}(X_i) + \mathrm{E}_{\mathrm{global}}(X), \quad (1)$$

where $X = X_1 \cup X_2$ is the label variables of all pixels in the input image pair, X_1 and X_2 denote the pixel label variables of image 1 and 2, respectively. For any

Figure 1. *Accuracy and speed of our method.*

$x_p \in X$, $x_p = 1$ indicates the corresponding pixel belonging to foreground, otherwise $x_p = 0$. The first term of (1) ensures co-segmentation to produce good foreground/background separation for image i individually ($i \in \{1, 2\}$), thus can be expressed as:

$$\mathrm{E}_{\mathrm{seg}}(X_i) = \sum_{p} w_{p,i} x_{p,i} + \sum_{p \sim q} w_{pq,i} |x_{p,i} - x_{q,i}|, \quad (2)$$

where $w_{p,i}$ encodes the cost of labeling pixel p as foreground in image i, $w_{pq,i}$ represents the cost of separating neighboring pixels p and q into different labels. The second term of (1) encourages the extracted foreground regions in image 1 and 2 are as similar as possible, which can be defined as the distance between their foreground un-normalized histograms:

$$\mathrm{E}_{\mathrm{global}}(X) = \alpha \sum_{k=1}^{K} \left(h_{k,1} - h_{k,2}\right)^2, \quad (3)$$

where $h_{k,i} = \sum_{p \in \mathcal{B}_{k,i}} x_{p,i}$ with $\mathcal{B}_{k,i}$ being the kth bin of image i, K denotes the number of bins, coefficient $\alpha \geq 0$ controls the relative importance of $\mathrm{E}_{\mathrm{global}}$.

Due to the existence of both submodular and supermodular terms, in most cases, the co-segmentation energy function (1) is NP-hard. An unpleasant result is that the complexity of pixel-level co-segmentation, e.g., [5], exponentially grows up for increasing image sizes. This significantly limits the application of pixel-level image co-segmentation in handling real-world high-resolution image pairs.

*Corresponding author. Tel: (+86) 22-2740-6538.

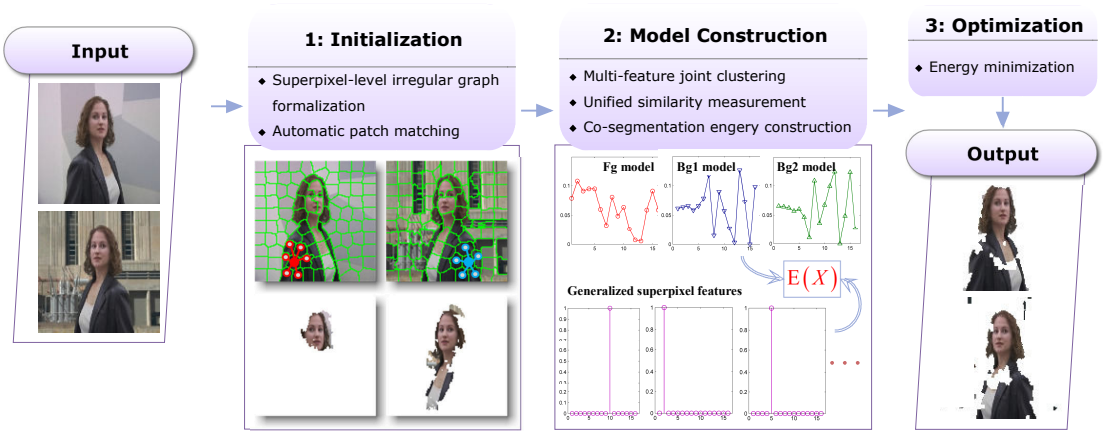

Figure 2. *The proposed method for scalable image co-segmentation using multimodal features.*

To this end, in this paper, we particularly study how to realize scalable co-segmentation without sacrificing labeling accuracy. We also interest in generalizing the original color histogram matching framework [9] to allow multi-features in co-segmentation energy function (1) for better performance. Specifically, we first present a general scalable framework for applying multimodal features in image co-segmentation based on superpixel-level irregular graph formulation and multi-feature joint clustering. Then, we propose a unified similarity metric for fusing pointwise color and regional covariance features. As shown in Fig. 1 and other experiments, the proposed method can achieve comparable or better segmentation accuracy, with superior scalability over state-of-the-art methods. We also find that multimodal features can always help to improve the accuracy of image co-segmentation.

2. Scalable Image Co-Segmentation

As shown in Fig. 2, a *general solution* to *scalable* and fully *automatic* image co-segmentation using multi-features should at least be composed of three key steps: (i) initialization; (ii) model construction; and (iii) optimization. For the purpose of good scalability, through initialization we first establish a *resolution-independent* representation $\mathcal{G}$ to the input image pair $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2\}$ and construct reasonable foreground/background initialization $X^{(0)}$ simultaneously. Then, in model construction, we design a unified similarity metric to embody the contributions of multiple features into the co-segmentation energy function (1). In the last step, we minimize the embodied energy function by an appropriate solver to efficiently obtain suboptimal labeling $\hat{X}$.

Noth that the accuracy of co-segmentation may be further improved by iteratively refining the foreground/background models with the current labeling $\hat{X}^{(t)}$ with t denoting the iteration number [11]. In this paper, however, we prefer *non-iterative* co-segmentation, i.e., $t = 1$, for good scalability.

2.1. Initialization

For a given image pair $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2\}$, we first obtain their SLIC superpixel representation [1], and construct an irregular graph formulation $\mathcal{G}$, see column 2 in Fig. 2 for an example, according to superpixels' spatially adjacent relationship. At the same time, we initialize the foreground/background labeling $X^{(0)} = \{X_1^{(0)}, X_2^{(0)}\}$ via automatic patch matching [2]. All subsequent computations of our method are based on $\mathcal{G}$ and $X^{(0)}$.

2.2. Co-segmentation with multi-features

To realize multi-feature co-segmentation, we need a general metric to measure the similarity or likelihood of a superpixel under the given foreground/background model. This similarity metric should also be able to take all features' contributions into account. For instance, for a given superpixel p, we use the mean color $\bar{\mathbf{c}}_p = \frac{1}{|\mathcal{S}_p|}\sum_{j\in\mathcal{S}_p}\mathbf{c}_j$ and regional covariance matrix $\mathbf{V}_p$ [10]

$$\mathbf{V}_p = \big(\mathbf{F}_p - \mu(\mathbf{F}_p)\big)^T\big(\mathbf{F}_p - \mu(\mathbf{F}_p)\big) \qquad (4)$$

as its multimodal features, where $\mathcal{S}_p$ is the set of pixels for superpixel p, $\mathbf{c}_j = [R_j, G_j, B_j]^T$ is the RGB color of pixel j, $|\mathcal{S}_p|$ denotes the number of pixels in p, $\mathbf{F}_p^T = [\mathbf{f}_{1,p}, \cdots, \mathbf{f}_{|\mathcal{S}_p|,p}]$ is the feature matrix of superpixel p,

$$\mathbf{f}_{j,p} = [x_j, y_j, R_j, G_j, B_j]^T \qquad (5)$$

is the 5 position and appearance features used in computing covariance features (4). Clearly, the sizes of feature matrix $\mathbf{F}_p$ and superpixel covariance matrix $\mathbf{V}_p$ are

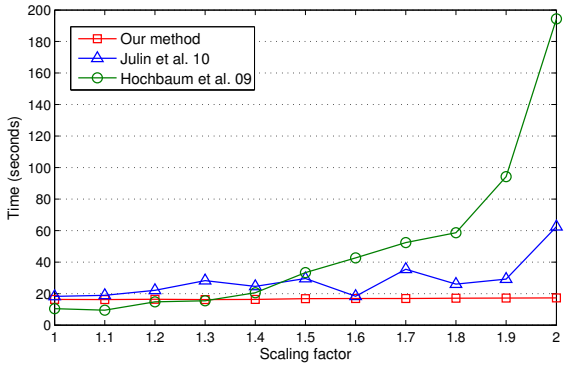

Figure 3. *Comparison of the proposed method and two state-of-the-art methods [5, 6] in algorithmic scalability.*

$|\mathcal{S}_p| \times 5$ and 5×5, respectively. For any two superpixels p and q, we compute its distance by

$$\mathrm{D}(p,q) = \lambda \|\bar{\mathbf{c}}_p - \bar{\mathbf{c}}_q\|_2 + (1-\lambda)\Big(\sum_{f=1}^{5} \ln^2 \rho_f(\mathbf{V}_p, \mathbf{V}_q)\Big)^{\frac{1}{2}}, \tag{6}$$

where $\|\cdot\|_2$ is the Euclidean distance, $\rho_f(\mathbf{V}_p, \mathbf{V}_q)$ is the fth generalized eigenvalues of $\mathbf{V}_p$ and $\mathbf{V}_q$ [10], parameter $0 \leq \lambda \leq 1$ is the weighting coefficient.

Instead of deriving individual color histograms of $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2\}$ [9], we jointly grouping all superpixels in both images into K clusters with the hybrid dissimilarity metric defined in (6). Using the K cluster centers as a common vocabulary $\mathcal{V} = \{\mathcal{C}_1, \cdots, \mathcal{C}_K\}$, the multi-features of superpixel p and the foreground model (i.e., a group of particular superpixels) can both be expressed in a unified manner as un-normalized histograms over $\mathcal{V}$, i.e., H_p and H_{fg} as shown in the 3rd column of Fig. 2. Then, we can measure the cost of labeling a superpixel p in image i as foreground:

$$w_{p,i} = \beta \|H_p, H_{\mathrm{fg}}\|_{\mathrm{emd}}, \tag{7}$$

where $\|\cdot\|_{\mathrm{emd}}$ indicates the EMD distance between two histograms [7], parameter $\beta \geq 0$ modulates the relative influence of labeling cost in (2). Besides, we apply the Potts model to encouraging spatial coherence [4], i.e.,

$$w_{pq,i} = \gamma, \tag{8}$$

where $\gamma \geq 0$ is a parameter penalizing inconsistent labeling in (2).

2.3. Optimization

From Eqs. (1)-(3), (7) and (8), we finally construct a general co-segmentation energy function with the capability of encoding multiple superpixel-level features. We empirically find that the proposed co-segmentation energy function can usually be effectively solved by belief propagation (BP) [12].

3. Experimental Results

In this section, we tested the proposed method on common datasets [9, 11], and compared its performance to three state-of-the-art methods [5, 6, 8] in terms of labeling accuracy and scalability. Note that, all results of the proposed method reported in this paper were obtained using a set of non-optimally-tuned parameters. Most results of the comparative methods were directly borrowed from their original papers [5, 6, 8]; while the others were generated using their original implementations with manually-tuned parameters by ourselves.

Labeling accuracy. We first evaluated segmentation accuracy. Table 1 shows the error rates for five typical testing image pairs of our method and three alternatives Alg.#1 [5], Alg.#2 [6], and Alg.#3 [8]. Note that, in Table 1, the error rate with bold font indicates the best result among all four testing methods for the corresponding image pair. We observe that our method is able to produce generally comparable and sometimes better labeling accuracy, as compared to the state-of-the-art. Note, for image pairs "bear" and "coke", our method produced the best co-segmentation results.

Fig. 1 and the last row of Fig. 4 demonstrate more results of our method. We can clearly see its capability of efficiently producing accurate co-segmentation results.

Table 1. *Co-segmentation error rates on typical testing image pairs of the proposed method and three representative algorithms, i.e., Alg.#1 [5], Alg.#2 [6], and Alg.#3 [8].*

	Alg.#1	Alg.#2	Alg.#3	Our method
stone	1.2%	**0.9%**	4.7%	1.7%
bear	3.9%	5.5%	8.6%	**2.8%**
dog	**3.5%**	6.4%	5.6%	5.7%
amira	4.5%	*16%*	**1.6%**	3%
coke	13.1%	*16.9%*	5%	**4.5%**

Scalability. We then compared the scalability of the proposed method with Alg.#1 [5] and Alg.#2 [6]. For this purpose, we selected two image pairs with different original sizes and different degrees of co-segmentation difficulty. We resampled the image pairs with 10 scaling factors from 110% to 200% with a step of 10%. For each testing method, 10 sets of randomly-generated parameters were used to evaluate its running time. Fig. 3 shows the average speed of all three testing methods on increasing scaling factors. We observe that both complexities of Alg.#1 [5] and Alg.#2 [6] increase rapidly

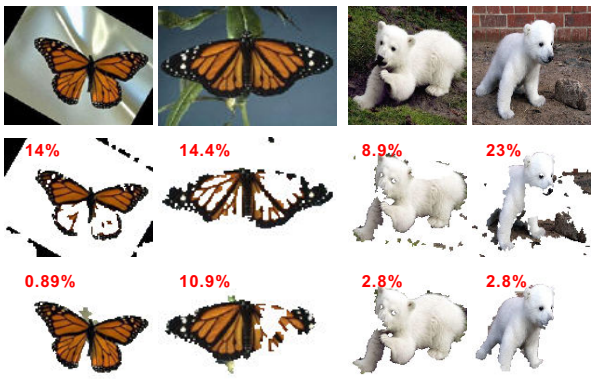

Figure 4. *Effectiveness of multi-feature co-segmentation. The 1st row is the input image pairs. Row 2 and 3 show the results of our method using color feature only, and using both color and region covariance features, respectively.*

as the resolution of input image pairs increases, while the speed of the proposed method is quite stable, thanks to the superpixel-level irregular graph formulation and unified similarity measurement. Note that, the complexity of Alg.#1 [5] increases much more quickly than the others. This is because it uses auxiliary variables in their optimization formulation. The increase of image resolution will both increase the number of pixel label variables and the number of additional auxiliary variables in Alg.#1 [5].

Effectiveness of multi-features. We are also interested in comparing the performance of the proposed method using single color feature and multi-features. Among the four main parameters $(\alpha, \beta, \gamma, \lambda)$ of our method, only λ is related to the weights of different features in the framework (6). Hence, for each image pair, we first set $\lambda = 1$ and quickly tuned α, β and γ using single color feature only. Once obtained reasonable results, we fixed α, β and γ, and then focused on tuning λ by gradually decreasing its value from 1. From our experiments, we find that: (i) for most image pairs our method using only color feature can produce good co-segmentation results; and more importantly (ii) using both color and regional covariance features can always produce better labeling accuracy than using color feature only with the same parameter-configuration of α, β and γ. Fig. 4 shows two representative examples of the superior effectiveness of multi-feature co-segmentation over single color-based co-segmentation.

4. Conclusions

This paper has introduced a general solution for superpixel-level scalable image co-segmentation using multimodal features. Besides, we have also particulary proposed a scalable co-segmentation algorithm using color and covariance features. Different from classical color-based co-segmentation [9, 11], to achieve both scalability and generality to multi-features, we have provided a unified similarity metric to embody the superpixel-based irregular graph representation and contributions of multiple features into a general energy function. Experimental results have validated the superior performance of our method in much better scalability over previous alternatives, and its ability to generate reasonable foreground labelings with comparable or even better accuracy. In the near future, we plan to explore the applications of more discriminative features, such as MSER, within the proposed framework.

Acknowledgement. This work is supported by the Program for New Century Excellent Talents in University (NCET-11-0365), the National Natural Science Foundation of China (61100121 and 61100122), and the research fund for Doctoral Program of Higher Education (20110032120036).

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels. In *EPFL Technical Report 149300*, 2010.

[2] C. Barnes, E. Shechtman, D. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In *ECCV*, 2010.

[3] J. Cech, J. Matas, and M. Perdoch. Efficient sequential correspondence selection by cosegmentation. *IEEE TPAMI*, 32(9):1568–1581, 2010.

[4] W. Feng, J. Jia, and Z.-Q. Liu. Self-validated labeling of Markov random fields for image segmentation. *IEEE TPAMI*, 32(10):1871–1887, 2010.

[5] D. S. Hochbaum and V. Singh. An efficient algorithm for co-segmentation. In *CVPR*, 2009.

[6] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

[7] H. Ling and K. Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE TPAMI*, 29(5):840–853, 2007.

[8] L. Mukherjee, V. Singh, and J. Peng. Scale invariant cosegmentation for image groups. In *CVPR*, 2011.

[9] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *CVPR*, 2006.

[10] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 2006.

[11] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation revisited: Models and optimization. In *ECCV*, 2010.

[12] J. Yedidia, W.-T. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS*, 2000.