From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model report

Paper's Authors

Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai

Advisor

Dr. Maryam Abedi

Student

Mohammad Shahpouri

October
2022

Contents

List of Figures

List of Tables

List of Equations

1 Co-segmentation energy function

An energy function is propsed for co-segmenting M images $\{I\}_{i=1}^{M}$. An over-segmentation technique [1] is applied to each image, and partition I_i into n_i superpixels. co-segmenting these M images is to find the binary labels $\{\mathbf{x}^i\}_{i=1}^{M}$ minimizing the following energy function:

$$\begin{aligned} F(\{\mathbf{x}^i\}) &= \sum_i L_i(\mathbf{x}^i) + \lambda \cdot E(\{\mathbf{x}^i\}) \\ &= \sum_i L_i(\mathbf{x}^i) + \lambda \sum_{i,j} G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j) \end{aligned} \tag{1}$$

where $L_i(x^i)$ is the *within-image* MRF energy of the labeling $\mathbf{x}^i$ on I^i, $G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j)$ is the *between-image* energy, and λ weighs the importance of the global energy term $E(\{\mathbf{x}^i\})$.

1.1 Co-saliency prior

SIFT feature [2], $\mathbf{g}_j^i$, is applied on every five pixels. For each $\mathbf{g}_j^i$ the *distance* to its most similar point on image I^k is calculated by:

$$d(\mathbf{g}_j^i, I^k) = \min_l \|\mathbf{g}_j^i - \mathbf{g}_l^k\| \tag{2}$$

$\mathbf{g}_j^i$ is now associated with $M-1$ distances $\{d(\mathbf{g}_j^i, I^k)\}_{k\neq i}$. To derive $\bar{d}_j^i$ average of the first half smallest distances are computed. Then, sigmoid function is utilized to define the weight w_j^i by:

$$w_j^i = \frac{1}{1 + \exp(-\frac{\mu - \bar{d}_j^i}{\sigma})} \tag{3}$$

where μ and σ are the parameters related to the shape of the sigmoid function. $\mu = 0.8$ and $\sigma = 0.2$

1.2 Within-image MRF energy

First the cost of labeling a superpixel $\mathbf{p}_j^i$ is computed as:

$$\alpha_j^i = \sum_{k \in \mathbf{p}_j^i} \tau - \tilde{s}_k^i \tag{4}$$

where τ is a parameter to be adjusted. $\tilde{s}_k^i$ is co-saliency map value of image I^i at pixel k. Second, calculating the cost of assigning different labels to two adjacent superpixels is:

$$\beta_{j,k}^i = \sum_{(l,m)\in B_{j,k}^i} \exp\left(-\frac{\|\mathbf{v}_l^i - \mathbf{v}_m^i\|^2}{2\sigma_{RGB}^2}\right) \tag{5}$$

where $\mathbf{v}_l^i$ and $\mathbf{v}_m^i$ are the respective RGB values of pixels l and m, and $B_{j,k}$ includes all the pairs of adjacent pixels across the boundary of superpixels $\mathbf{p}_j^i$ and $\mathbf{p}_k^i$. With Equation 4 and Equation 5, the exact form of $L_i(\mathbf{x}^i)$ can then be stated as follows:

$$L_i(\mathbf{x}^i) = \sum_{j=1}^{n_i} \alpha_j^i x_j^i + \sum_{(j,k)\in\mathcal{E}^i} \beta_{j,k}^i \delta[x_j^i \neq x_k^i] \tag{6}$$

where n_i is the total number of superpixels in I^i, δ is an indicator function that outputs 1 when the statement is true. $\beta_{j,k}^i > 0$ for all $(j,k) \in \mathcal{E}^i$ ensures the following important regularity about $L_i(\mathbf{x}^i)$.

Property 1 *The within-image MRF energy $L_i(\mathbf{x}^i)$ defined in Equation 6 is submodular.*

1.3 Global energy term

To compute global energy term each superpixel is represented by an unnormalized histogram $\mathbf{h}$.

$$\mathbf{H}_f^i = \sum_{k=1}^{n_i} \mathbf{h}_k^i x_k^i \quad \text{and} \quad \mathbf{H}_b^i = \sum_{k=1}^{n_i} \mathbf{h}_k^i (1 - x_k^i) \tag{7}$$

Then, the histogram of I^i is denoted as:

$$\mathbf{H}^i = \sum_{k=1}^{n_i} \mathbf{h}_k^i = \mathbf{H}_f^i + \mathbf{H}_b^i \tag{8}$$

Between-image energy $G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j)$ can be calculated as:

$$G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j) = \|\mathbf{H}_f^i - \mathbf{H}_f^j\|_2^2 - \sum_{k\in\{i,j\}} c_1^k \|\mathbf{H}_f^k - c_2^k \mathbf{H}_b^k\|_2^2 \tag{9}$$

where c_1^* decides the influence of the dissimilarity, and c_2^* is to balance the foreground and the background histograms.

By substituting $\mathbf{H}_b^i = \mathbf{H}^i - \mathbf{H}_f^i$ into Equation 9, and taking the definition of $\mathbf{H}_f^i$ in Equation 7, we obtain:

$$\begin{aligned} G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j) = C - 2\sum_{l,m} \left\langle \mathbf{h}_l^i, \mathbf{h}_m^j \right\rangle x_l^i x_m^j + \\ 2c_1 c_2 (1 + c_2) \times \sum_{k\in\{i,j\}} \sum_{l=1}^{n_k} \left\langle \mathbf{h}_l^k, \mathbf{H}^k \right\rangle x_l^k + \\ (1 - c_1(1 + c_2)^2) \times \sum_{k\in\{i,j\}} \sum_{l,m} \left\langle \mathbf{h}_l^k, \mathbf{h}_m^k \right\rangle x_l^k x_m^k \end{aligned} \tag{10}$$

where C is a constant term. $c_1 = \frac{1}{(1+c_2)^2}$. Finally, by setting $c = \frac{c_2}{1+c_2}$, $G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j)$ becomes:

$$C - 2\sum_{l,m} \left\langle \mathbf{h}_l^i, \mathbf{h}_m^j \right\rangle x_l^i x_m^j + 2c \times \sum_{k\in\{i,j\}} \sum_{l=1}^{n_k} \left\langle \mathbf{h}_l^k, \mathbf{H}^k \right\rangle x_l^k \tag{11}$$

Property 2 *The total energy function F defined in Equation 9 is submodular, and hence the proposed energy minimization can be optimally solved by the graph-cut algorithm.*

2 Learning visual vocabulary

Suppose that J pixels are uniformly sampled from each image, and represent each pixel by a SIFT feature vector $\mathbf{z}$. To cluster all these pixels over $\{I^i\}_{i=1}^M$ into K visual words, an assignment table A of size $M \times J \times K$, and the following optimization problem are considered:

$$\begin{aligned} \min_{\{\mu_k\}_{k=1}^K, A} \sum_{k=1}^{K}\sum_{i=1}^{M}\sum_{j=1}^{J}(\|\mathbf{z}_{i,j} - \mu_k\| \cdot A_{i,j,k}) + \\ \eta \times \sum_{k=1}^{K} \sqrt{\sum_{i=1}^{M}\left(\sum_{j\in R^i}^{J} A_{i,j,k}\right)^2} \\ \text{subject to} \quad A_{i,j,k} \in \{0,1\}, \\ \sum_k A_{i,j,k} = 1, \forall i,j \end{aligned} \tag{12}$$

where $\{\mu_k\}$ are the cluster centers and $\eta = 4$ controls the influence of the regularization term.

3 Experimental results

Weizman horses, MSRC database, and gnome dataset are used to evaluate the performance of the method.

3.1 Quantitative result

Table 1. Co-segmentation accuracy. The results by the proposed method, measured in the pixel accuracy, are reported in the rightmost seven columns. When the global energy term E in Equation 1 is included, visual words can be obtained either by K-means or by K-means with $L_{1,2}$ regularization.

Dataset	Num. of images	DC [3]	Without global term		K-means		K-means + $L_{1,2}$		
			Saliency	Co-Saliency	Saliency	Co-Saliency	Saliency	Co-Saliency	$\{c_2, \tau\}$
Cars front	6	87.65%	77.01%	79.01%	83.27%	88.50%	88.04%	**90.78%**	90.46%
Cars back	6	85.10%	76.22%	77.63%	79.72%	88.50%	85.34%	**85.76%**	**85.76%**
Bike	30	63.30%	70.90%	72.38%	75.06%	76.67%	75.52%	**76.76%**	76.60%
Cat	24	74.40%	83.06%	79.80%	85.78%	86.36%	86.34%	**86.68%**	**86.68%**
Plane	30	75.90%	85.91%	86.22%	86.58%	86.80%	86.92%	**87.66%**	87.21%
Face	30	84.30%	78.54%	78.96%	84.41%	85.51%	85.08%	**87.27%**	85.76%
Cow	30	81.60%	88.40%	88.71%	91.25%	91.30%	91.10%	**91.36%**	90.92%
Horse	30	80.10%	78.72%	76.59%	85.30%	86.00%	85.57%	**86.36%**	84.36%
Gnome	4		89.29%	93.56%	93.28%	95.21%	95.00%	**95.29%**	95.12%

3.2 Qualitative result

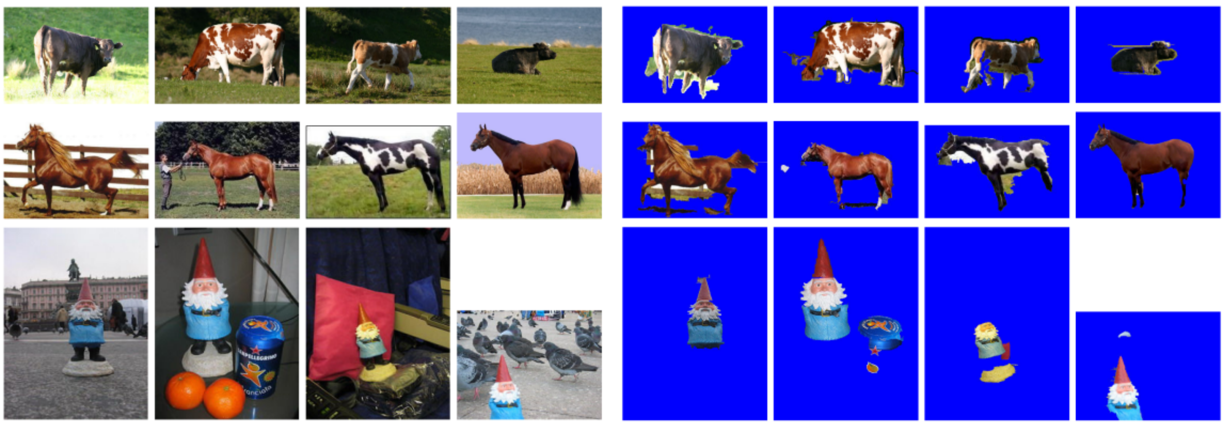

Figure 1. Examples of the input images and the co-segmentation results.

References

[1] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *International Journal of Computer Vision*, vol. 59, pp. 167–181, Sep 2004. 1

[2] D. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 vol.2, 1999. 1

[3] A. Joulin, F. Bach, and J. Ponce, “Discriminative clustering for image co-segmentation,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1943–1950, 2010. 4