From Co-saliency to Co-segmentation: An Efficient and Fully Unsupervised Energy Minimization Model

Kai-Yueh Chang[1,2] Tyng-Luh Liu[1] Shang-Hong Lai[2]
[1]Institute of Information Science, Academia Sinica, Taiwan
[2]Department of Computer Science, National Tsing Hua University, Taiwan

Abstract

We address two key issues of co-segmentation over multiple images. The first is whether a pure unsupervised algorithm can satisfactorily solve this problem. Without the user's guidance, segmenting the foregrounds implied by the common object is quite a challenging task, especially when substantial variations in the object's appearance, shape, and scale are allowed. The second issue concerns the efficiency if the technique can lead to practical uses. With these in mind, we establish an MRF optimization model that has an energy function with nice properties and can be shown to effectively resolve the two difficulties. Specifically, instead of relying on the user inputs, our approach introduces a co-saliency prior as the hint about possible foreground locations, and uses it to construct the MRF data terms. To complete the optimization framework, we include a novel global term that is more appropriate to co-segmentation, and results in a submodular energy function. The proposed model can thus be optimally solved by graph cuts. We demonstrate these advantages by testing our method on several benchmark datasets.

1. Introduction

Figure-ground segmentation has long been a challenging problem in computer vision. Apart from the difficulties in establishing an effective framework to divide the image pixels into *meaningful* groups, the notions of figure and ground often need to be properly defined by providing either user inputs, *e.g*., [10, 25, 28] or object models, *e.g*., [3, 19]. The idea of co-segmentation, first introduced by Rother *et al*. [26], is another possibility to more explicitly cast the problem as simultaneously segmenting an image pair to locate their *common* foreground object. However, as is illustrated in Figure 1, the ambiguity of what the foregrounds should be may still exist and can confuse a fully automated approach to correctly segment the desired object. On the other hand, by including more images containing the common object, the co-segmentation formulation can become more well-defined, but generally at the price of increasing the appearance variations of the foregrounds as well as the complexity to accomplish the task. The main focus of this paper is to tackle the aforementioned difficulties in performing co-segmentation over two or more images.

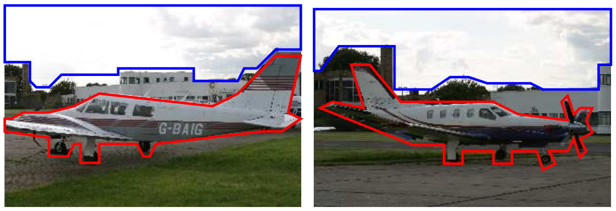


Figure 1. While humans would easily identify `airplane` (enclosing with red boundary) is the common object between the two images, the more consistent appearances in the `sky` areas (enclosing with blue boundary) may cause an automatic segmentation system to choose them as the more plausible foregrounds.

Specifically, we aim to address two key issues of co-segmentation: 1) Can a fully unsupervised approach satisfactorily solve the problem? 2) Is there a general energy minimization model to realize the co-segmentation process with efficiency? In the first issue, one practical and often-encountered scenario is that the appearance variations of the common object are more substantial than those in some areas of the backgrounds. (See Figure 1.) To overcome the dilemma, most of the existing techniques [1, 6, 11, 21] adopt an interactive scheme to guide the grouping process through ambiguous situations. We instead consider a *co-saliency* prior in the proposed energy function, and establish an unsupervised co-segmentation algorithm. The concept of saliency has been extensively studied by psychologists [14, 24, 27, 30]. Roughly speaking, it is related to the areas that most people may focus on when seeing a view. For co-segmentation, we are motivated to define a co-saliency model to generate image regions that are *similar* to each other across images, and meanwhile retain their distinctness within each image. To that end, we adopt the saliency results of Goferman *et al*. [8], and then filter out the areas that infrequently appear in most images. In the second issue, we observe that an important factor in solving the co-segmentation problem is the establishment of a suitable consistency measure between the foreground areas from any two different images. For those techniques based on MRF, the consistency check is often carried out by introducing a global term in the energy function, *e.g.*, [11, 21, 26]. Such a global term is typically defined by the similarity of two histograms from two potential foreground areas. Since the number of possible foreground regions in an image is two to the power of the number of superpixels (or pixels, in the extreme case), evaluating the global term over all the possible foreground pairs becomes pivotal concerning efficiency. This aspect of consideration is even more critical when more images are included for co-segmentation. In our approach, we have proposed a new and effective global term that satisfies the *submodular* condition [17]. The resulting energy minimization can thus be optimally solved by the graph-cut algorithm [4].

Another aspect of our effort is to investigate how to come up with a good feature representation for co-segmentation. In particular, we focus on the global energy term for the reason just described. As in our formulation the potential foreground regions evaluated by the global term are represented as histograms of *visual words*, it is constructive to explore whether enforcing the clustering criterion to consider the property that the images share a common object would result in a more effective *vocabulary*. That is, we may prefer that pixels (or sampled pixels) are grouped to form a visual word owing to not only having similar descriptor values but also spreading over *different* images. This point will be discussed in detail in Section 4. We summarize the main contributions of this paper as follows:

- Introduce a co-saliency prior to make the unsupervised co-segmentation possible.
- Establish a new global energy term to effectively solve co-segmentation over multiple images.
- Propose a useful regularization term in K-means objective function to encourage gathering pixels with similar appearance across different images.

2. Related work

Co-segmentation techniques most relevant to ours are those heavily relying on the regularity of a global energy term in their MRF model. Rother *et al*. [26] and Mukherjee *et al*. [21] respectively use the L_1 and the L_2 norm to measure the dissimilarity between foreground histograms. The main drawback of both is that solving the whole model becomes NP-hard. Hochbaum and Singh [11] subsequently propose a “reward” model that satisfies the submodular condition and therefore can be efficiently solved by graph cuts. However, the inner product of two *unnormalized* histograms representing the reward model is hard to give a meaningful explanation of why it yields a suitable similarity measure. Namely, a large inner product value by their model does not imply that the two unnormalized histograms are more similar. In [29], without directly comparing two histograms, Vicente *et al*. propose a new global model to favor a co-segmentation result that all pixels associated with a visual word are either uniformly from background or foreground. The criterion is reasonable when the desired foregrounds of both images are indeed instances of the same object with possible scale changes. It is not clear if the model can be extended to handle more challenging backgrounds, viewpoint changes, and appearance variations pertaining to the foreground object. While the above methods [11, 21, 26, 29] all include a global term in their MRF model and test on co-segmentation with only two images containing identical or similar objects, Joulin *et al*. [15] consider co-segmenting more than two images with different instances from a more general concept of the *same* object class. Their formulation treats co-segmentation as a two-cluster problem, and yields impressive results. However, since the goodness of clustering depends on the accuracy in evaluating the similarity between every two local patches (or superpixels), the framework seems to require fine over-segmentation, say, around 500 superpixels per image to give satisfactory performances, and therefore results in a less efficient implementation.

As is mentioned earlier, a number of co-segmentation methods need user inputs to facilitate the process. The approaches by Mukherjee *et al*. [21] as well as by Hochbaum and Singh [11] both require providing some scribbles (similar to those in GrabCut [25]). Instead of suggesting the scribbles at first, Batra *et al*. [1] propose to guide the user to input additional strokes on the area that is the hardest to decide the pixel labels. Without relying on the scribble cues, Cui *et al*. [6] assume that one of the images is hand-segmented. Rother *et al*. [26] add a constant penalty for assuming the background label to avoid the trivial solution that all pixels are labeled as background. Joulin *et al*. [15] divide pixels into two clusters, and let the user choose which cluster is more likely to be the common object cluster.

In passing, we notice that more recently Chen [5] has proposed a scheme to find the common salient objects between a pair of images by enhancing the similar and pre-attentive patches. However, it appears to be hard to generalize the formulation to the case of handling more than two images. Also, Rahtu *et al*. [22] and Ramanathan *et al*. [23] both take account of the saliency information in segmenting meaningful objects from a *single* image. Direct and feasible extensions of their approach to co-segmentation of two or more images are not obvious in view of the difficulty in sifting the saliency information from each image.

3. Co-segmentation energy function

Given a set of M images $\{I^i\}_{i=1}^{M}$ for co-segmentation, the foreground of I^i is simply the area containing an instance of the common object. We apply the over-segmentation technique provided in [7] to each image, and partition I^i into n_i superpixels. Then, the foreground and the background of I^i can be approximately represented by a binary label vector $\mathbf{x}^i \in \{0,1\}^{n_i}$. It follows that in the context of the proposed MRF model, co-segmenting these M images is to find the binary labels $\{\mathbf{x}^i\}_{i=1}^{M}$ minimizing the following energy function:

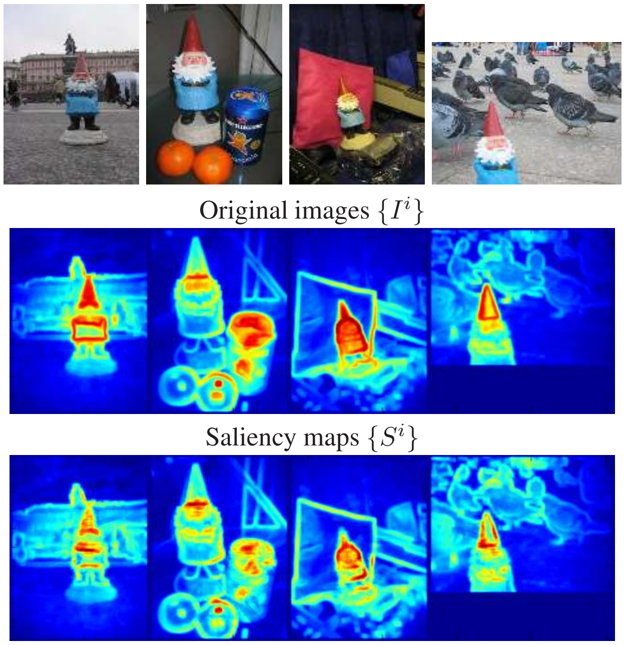
$$\begin{aligned} F(\{\mathbf{x}^i\}) &= \sum_i L_i(\mathbf{x}^i) + \lambda \cdot E(\{\mathbf{x}^i\}) \\ &= \sum_i L_i(\mathbf{x}^i) + \lambda \sum_{i,j} G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j) \quad (1) \end{aligned}$$

where $L_i(\mathbf{x}^i)$ 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 measuring the inconsistency between I^i and I^j under the labelings $\mathbf{x}^i$ and $\mathbf{x}^j$, and λ weighs the importance of the global energy term $E(\{\mathbf{x}^i\})$. In what follows, we shall first explain how the co-saliency prior is derived, and how to use the information to construct $\{L_i\}$. We then focus on the details of G and its useful property to complete the proposed energy minimization model.

3.1. Co-saliency prior

Saliency detection is often formulated as the search of the distinct areas in an image, *e.g*., [8, 12, 13] since human eyes are easily attracted by the unusual things with respect to the whole view. Our co-segmentation model assumes that in most of the images $\{I^i\}_{i=1}^{M}$, their detected salient areas contain at least parts of the foreground object. The assumption is reasonable as the object of interest usually has some distinguishable appearances from the rest and draw one's attention. This will be termed as the *distinctness* property within an image. On the other hand, the *repeatedness* property among images is also important for co-segmentation. Namely, we prefer that a salient area in an image can be repeatedly detected in others. Based on these observations, we consider the single-view saliency model of Goferman *et al*. [8], and concentrate on those parts of saliency maps that frequently repeat in most images, *i.e*.,

$$\text{Co-saliency} = \text{Saliency} \times \text{Repeatedness}.$$



Original images $\{I^i\}$

Saliency maps $\{S^i\}$

Co-saliency maps $\{\tilde{S}^i\}$

Figure 2. Saliency versus co-saliency. Within the second image to the left, the can is salient. However, most pixels of that area have less repeatedness weights due to that the can does not appear in other images. It follows that the corresponding co-saliency map would have smaller co-saliency values around that area.

Let S^i and $\tilde{S}^i$ be the saliency and the co-saliency maps of I^i, and their value at pixel j is denoted as s^i_j and $\tilde{s}^i_j$, respectively. To obtain the co-saliency map $\tilde{S}^i$, we adjust each s^i_j by multiplying a weight w^i_j that can be thought as the likelihood of repeatedness over $\{I^k\}_{k\neq i}$. More specifically, we focus on those image pixels whose saliency value is larger than $0.6 \times s^i_{\max}$ where $s^i_{\max}$ is the maximal saliency value of I^i. And in such distinct areas, we sample a point every five pixels and describe it by a SIFT feature [20]. Let $\mathbf{g}^i_j$ be the SIFT feature of point j on image I^i. For each $\mathbf{g}^i_j$, we compute the *distance* to its most similar point on image I^k by

$$d(\mathbf{g}^i_j, I^k) = \min_l \|\mathbf{g}^i_j - \mathbf{g}^k_l\|. \quad (2)$$

Thus, according to (2), each $\mathbf{g}^i_j$ is now associated with $M-1$ distances $\{d(\mathbf{g}^i_j, I^k)\}_{k\neq i}$. We then average the first half smallest distances to derive $\bar{d}^i_j$, and use the sigmoid function to define the weight w^i_j by

$$w^i_j = \frac{1}{1+\exp\left(-\frac{\mu-\bar{d}^i_j}{\sigma}\right)} \quad (3)$$

where μ and σ are the parameters related to the shape of the sigmoid function. ($\mu = 0.8$ and $\sigma = 0.2$ in all our experiments.) While the above procedure yields only the weights of the sampled points, the weights of all the remaining pixels of I^i can be approximated by interpolation from those of their closest sampled points. Finally, we rescale the co-saliency values of $\tilde{S}^i$ to $[0,1]$. (See Figure 2.)

3.2. Within-image MRF energy

We are now ready to define the within-image energy $L_i(\mathbf{x}^i)$ in (1) of binary labeling $\mathbf{x}^i$ over superpixels $\{\mathbf{p}_j^i\}$ of I^i. Like in most of the conventional MRF models, L_i contains a data term and a pairwise *smoothness* term. To specify the two terms, we need two additional definitions. The first pertains to the cost of labeling a superpixel, say, $\mathbf{p}_j^i$ as foreground, and is given by

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

where τ is a parameter to be adjusted and its discussions will be provided in Section 5.1. The second definition concerns the cost of assigning different labels to two adjacent superpixels. Let $\mathcal{E}^i$ be the edge set that encodes the adjacency relations of $\{\mathbf{p}_j^i\}$ and $\beta_{j,k}^i$ be the cost of different labels between $\mathbf{p}_j^i$ and $\mathbf{p}_k^i$, $(j,k) \in \mathcal{E}^i$. In particular, we have

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

where $\mathbf{v}_l^i$ and $\mathbf{v}_m^i$ are the respective RGB values of pixels l and m, and $B_{j,k}^i$ includes all the pairs of adjacent pixels across the boundary of superpixels $\mathbf{p}_j^i$ and $\mathbf{p}_k^i$. (In our implementation σ_{RGB} is set to $20/256$.) With (4) and (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, and δ is an indicator function that outputs 1 when the statement is true. The fact that $\beta_{j,k}^i > 0$ 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 (6) is submodular.*

3.3. Global energy term

In evaluating the global energy term $E(\{\mathbf{x}^i\})$ in (1), like [11, 21, 26], we represent each superpixel by an unnormalized histogram $\mathbf{h}$. It follows that the summation of the histograms of all the superpixels within an area also forms this area's representation. Given a binary labeling $\mathbf{x}^i$ over image I^i, the implied foreground and background can be respectively represented by

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

We further denote the histogram of I^i as

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

From (1), establishing the global term can be reduced to specifying the between-image energy $G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j)$. We observe that good co-segmentation results often share two important attributes—not only the foregrounds are similar to each other but also each of them should be dissimilar to its respective background. We thus define

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

where c_1^* decides the influence of the dissimilarity, and c_2^* is to balance the foreground and the background histograms; otherwise, directly comparing these two un-normalized histograms may not be reasonable, since their corresponding areas can be of very different sizes. Note that the dissimilarity measure in (9) is between the entire foreground and background areas, which is different from the pairwise terms of L^i in (6) measuring only the local dissimilarities between superpixels. For simplicity, we assume hereafter c_1^* and c_2^* are respectively set to the same values c_1 and c_2.

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

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

where C is a constant term. Indeed the first three terms in the RHS of (10) satisfy the submodular condition. Whether G is a submodular function only depends on if the coefficient $1 - c_1(1+c_2)^2$ of the last term is not greater than 0. We let $c_1 = \frac{1}{(1+c_2)^2}$ so that G can be submodular, and meanwhile assume a simpler form. Finally, by setting $c = \frac{c_2}{1+c_2}$, $G(\mathbf{x}^i, \mathbf{x}^j, I^i, I^j)$ becomes

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

From (11), we find that the global energy term in Hochbaum and Singh [11] is a special case of our model when $c = 0$ (*i.e.*, $c_2 = 0$). On the other hand, when c is set

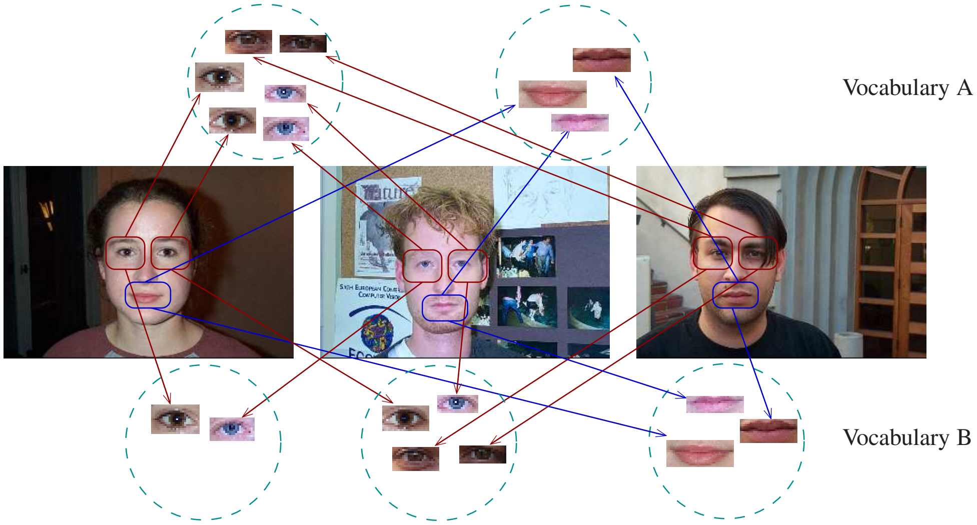


Figure 3. Two examples of visual vocabularies. K-means clustering with the proposed regularization term would more likely yield vocabulary A, which, compared with vocabulary B, is more compact and has two representative visual words (clusters).

to 1, it implies that c_1 is close to 0, and the proposed model is close to that based on L_2 norm used in [21]. Indeed, by introducing only one extra parameter in our approach, we are able to establish a more appropriate global energy term, and effectively tackle co-segmentation over more than two images, which is a much more complicated problem than co-segmenting an image pair.

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

4. Learning visual vocabulary

Our discussions in the previous section point out that the histogram representation of a superpixel plays an important role in formulating the global energy term. Since such a histogram is simply composed of the frequencies of the visual words [18], how these words are derived should, in turn, be a key factor. In our implementation, we have considered two different schemes. The first is standard that the visual words (clusters) are obtained by carrying out K-means clustering over sampled pixels, and it indeed gives satisfactory implementation of our method. In addition, we also explore the assumption that the images contain instances of a common object in the clustering process. Suppose for the moment that in each image I^i we have a region R^i intersecting the underlying foreground. (R^i will be discussed in Section 5.2.) We then consider devising a clustering scheme to prefer: 1) across-image, similar pixels of $\{R^i\}$ assigned to a same cluster should come from as many different images as possible, and 2) within-image, similar pixels of R^i should fall into the same cluster. Clearly, the set of visual words containing pixels from $\{R^i\}$ would become more *representative* and *compact*. (See Figure 3 for an illustration.)

So, the other clustering scheme used in our implementation is to take account of the above two useful properties. To this end, we add a regularization term in the K-means objective function by utilizing the effects of L_1 and L_2 norm. As described in [2], an L_1-norm regularization term tends to concentrate values on several entries of a vector while an L_2-norm regularization term instead spreads values over whole entries. Suppose that we uniformly sample J pixels 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, we consider an assignment table A of size $M \times J \times K$, and the following optimization problem:

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

where $\{\boldsymbol{\mu}_k\}$ are the cluster centers and η controls the influence of the regularization term. ($\eta = 4$ in all our experiments.) The justification of the regularization term in (12) can be best understood by first marginalizing A over R^i to obtain an $M \times K$ matrix. Now, each column of the matrix records the frequencies of a specific visual word (cluster) appearing in the M images. While the L_2 norm signaled by taking a square root is to make this visual word spread over all images, the L_1 norm implied by the outmost summation is to derive a compact set of visual words. For convenience, we will call this an $L_{1,2}$ regularization term.

Dataset	Num. of images	[15]	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%	81.86%	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%

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

Analogous to K-means clustering, we adopt an EM procedure to find $\{\boldsymbol{\mu}_k\}_{k=1}^K$ and A in (12). In E-step, we first relax $A_{i,j,k}$ to $[0,1]$ so that an NP-hard problem can be transformed into a convex optimization problem. We then use the `cvx` toolbox [9] to solve A, and discretize its entries to $\{0,1\}$ by setting $A_{i,j,k}$ to 1 if $k = \arg\max_l A_{i,j,l}$ and 0, otherwise. In M-step, we compute $\boldsymbol{\mu}_k$ by the mean of the feature vectors of the pixels assigned to cluster k.

5. Experimental results

For the sake of comparison, we test our method with the challenging datasets used in [15] which contain the Weizman horses and MSRC database. We also include the `Gnome` dataset in our experiments as it contains images with large illumination and viewpoint changes of the same object. Figure 4 shows some examples of these images and the co-segmentation results. (Note that the images in Weizman horses are resized to have the same larger dimension.)

5.1. Parameters

Our model has three parameters, $\{\lambda, \tau, c\}$. τ appears in (4), and its value is decided by running our algorithm without the global term. λ and c are introduced in (1) and (11), respectively. Recall that $c = c_2/(1 + c_2)$ and from (9), c_2 can be thought of the ratio of the average area of foregrounds to the average area of backgrounds over $\{I_i\}_{i=1}^M$. For each dataset, we uniformly sample c_2 from a given range, adjust λ heuristically, and then report the best result. Indeed, the set of parameters can be reduced to $\{\tau, c\}$ with slight decrease in accuracy, as λ can be tuned *unsupervisedly* by checking whether the co-segmentation results match the foreground-background ratio implied by c_2.

5.2. Accuracy

We test our co-segmentation method in seven different settings. First, the within-image energy $L_i(\mathbf{x}^i)$ can be implemented with the cost of assigning a foreground label according to either the co-saliency prior used in (4) or the saliency prior by replacing $\tilde{s}_k^i$ with s_k^i. Second, to single out the effect of the global term E in (1), the experiments are also performed with or without this global term. In case that E is included, the histogram representation will be used to describe a superpixel or an area of superpixels, and we further consider the two ways of constructing visual vocabularies described in Section 4. When the co-saliency prior and the $L_{1,2}$ regularization are used, we additionally test our method by tuning only τ and c. The respective results are reported in the rightmost seven columns of Table 1.

It can be inferred from the results in Table 1 that the co-saliency prior tends to yield better co-segmentation performances than the saliency prior, except implementing our model without using the global term to test the two datasets, `Cat` and `Horse`. And, such few exceptions are expected since the between-image factors are not considered here.

We next look into the importance of using the global energy term E in co-segmentation. In Table 1, the results by including E are those in the rightmost five columns, and they are uniformly superior to those without using the global term. However, considering the global term means the necessity of the two parameters λ and c, where the former controls its contribution, and the latter enables our model to tackle the high complexity of co-segmentation over more than two images.

The last factor discussed here that has a bearing on the co-segmentation accuracy is how the visual words are obtained. Recall that in Section 4, the $L_{1,2}$ regularization term is formulated based on the assumption that we have a region R^i that has a higher probability of intersecting the underlying foreground in image I^i. In practice, we have no access to such knowledge in an unsupervised approach. Nevertheless, a reasonable way to yield R^i is as follows. Besides the co-saliency map $\tilde{S}^i$, we also apply the *Gaussian center* prior [16] to I^i and generate, say, O^i. The superpixels intersecting the areas with the top 20% values of $\tilde{S}^i \times O^i$ are then included in R^i. The strategy is general in the sense that the ratio between the area of the resulting R^i to the area of I^i

Dataset	Avg. superpixels	Total superpixels	Time (in second)
Cars_front	96.83	581	0.65
Cars_back	104	624	0.85
Bike	96.9	2907	215.40
Cat	75.71	1817	19.84
Plane	90.67	2720	62.77
Face	87.73	2632	110.88
Cow	46.03	1381	6.96
Horse	94.03	2821	150.18
Gnome	77.50	310	0.22

Table 2. Inference time for each dataset.

can range from 35.70% to 68.56% in our test datasets. More importantly, it can reduce the unexpected effect of applying the $L_{1,2}$ regularization to the backgrounds, as is justified by the improved accuracy in the rightmost three columns.

5.3. Time complexity

We run our algorithm on a PC with Intel i7 CPU @ 2.8 GHz. In Table 2, the inference time in optimizing $\{\mathbf{x}^i\}$ with the energy defined in (1) is reported for each dataset. Compared with the technique of Joulin *et al*. [15], where the average number of superpixels is around 500 per image, and the inference time is about 8 minutes for a pair of images and 4 to 9 hours for 30 images, our method is more efficient. In particular, the proposed co-segmentation approach does not require over-segmenting each image into large number of superpixels, and can efficiently accomplish the task via an optimal labeling derived by the graph-cut algorithm.

6. Conclusion

Our main contribution is to introduce a new energy minimization model that is general enough to deal with the high complexity of simultaneously segmenting multiple images, and meanwhile, can be efficiently and optimally solved. In addition, we have proposed a useful regularization term for K-means clustering in learning the visual words for co-segmentation. If the inclusion of the three parameters (or two, for slight decreases in the co-segmentation accuracy) in our method is not considered, we have come close to establish a fully-unsupervised algorithm. Still, there are several issues remained to be explored. In particular, we would make efforts to further improve the quality of the saliency and co-saliency detection, extend co-segmentation to a more general concept, and bring in new applications.

Acknowledgements

This work is supported in part by NSC grants 97-2221-E-001-019-MY3 and 99-2221-E-001-011-MY3.

References

[1] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, pages 3169–3176, 2010.

[2] S. Bengio, F. Pereira, Y. Singer, and D. Strelow. Group sparse coding. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *NIPS 22*, pages 82–89. 2009.

[3] E. Borenstein and S. Ullman. Combined top-down/bottom-up segmentation. *PAMI*, 30(12):2109–2125, December 2008.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, November 2001.

[5] H. Chen. Preattentive co-saliency detection. In *ICIP*, 2010.

[6] J. Cui, Q. Yang, F. Wen, Q. Wu, C. Zhang, L. van Gool, and X. Tang. Transductive object cutout. In *CVPR*, pages 1–8, 2008.

[7] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, September 2004.

[8] S. Goferman, L. Zelnik Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, pages 2376–2383, 2010.

[9] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, Oct. 2010.

[10] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *CVPR*, pages 3129–3136, 2010.

[11] D. Hochbaum and V. Singh. An efficient algorithm for co-segmentation. In *ICCV*, pages 269–276, 2009.

[12] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8, 2007.

[13] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

[14] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203, March 2001.

[15] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950, 2010.

[16] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009.

[17] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 26(2):147–159, February 2004.

[18] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, June 2001.

[19] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. *IJCV*, 81(1), January 2009.

[20] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[21] L. Mukherjee, V. Singh, and C. Dyer. Half-integrality based algorithms for cosegmentation of images. In *CVPR*, pages 2028–2035, 2009.

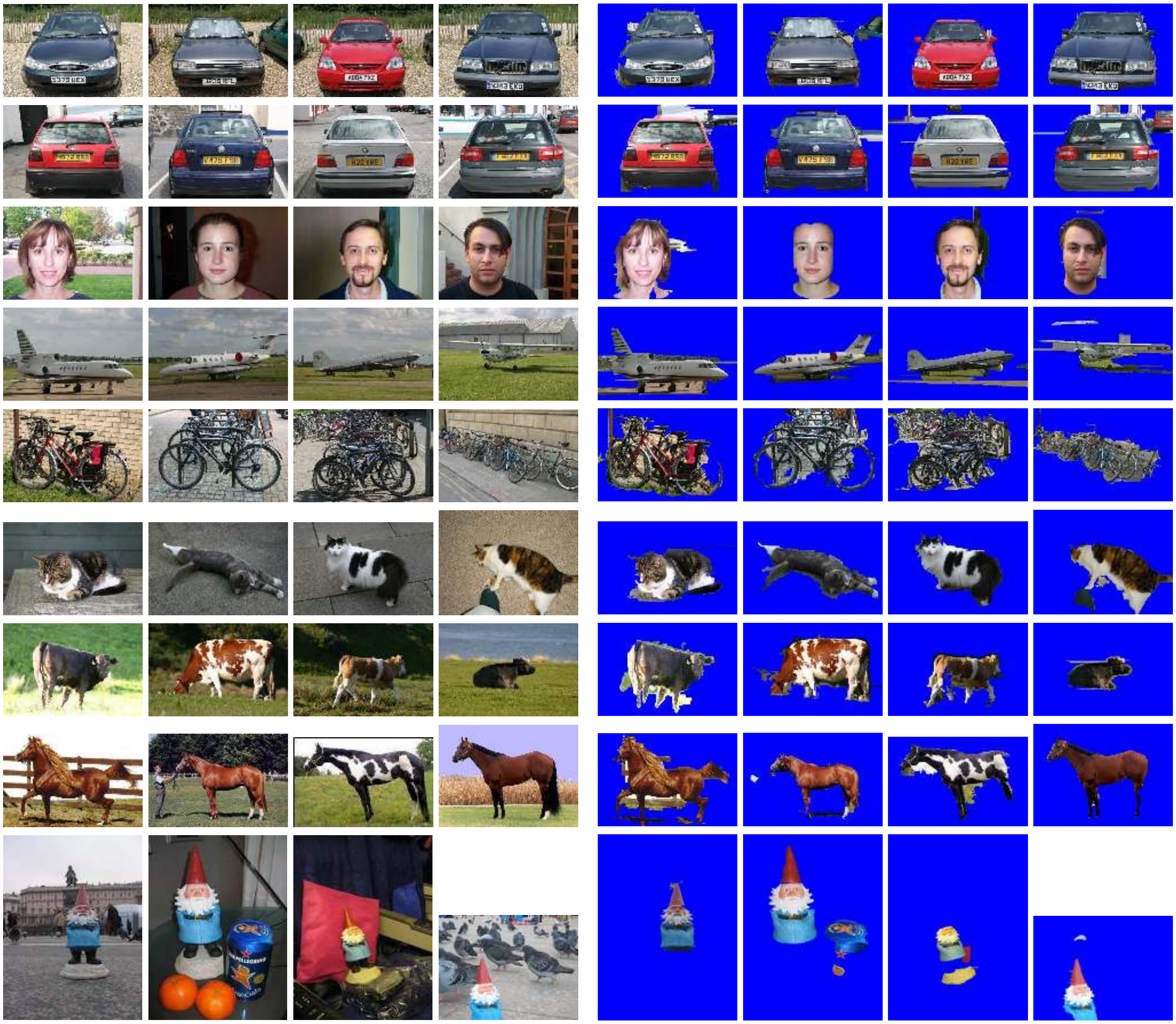


Figure 4. Examples of the input images and the co-segmentation results.

[22] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila. Segmenting salient objects from images and videos. In *ECCV*, pages V: 366–379, 2010.

[23] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T. Chua. An eye fixation database for saliency detection in images. In *ECCV*, pages IV: 30–43, 2010.

[24] R. Rao and D. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, January 1999.

[25] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[26] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching: Incorporating a global constraint into MRFs. In *CVPR*, pages I: 993–1000, 2006.

[27] A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, January 1980.

[28] O. Veksler. Star shape prior for graph-cut image segmentation. In *ECCV*, pages III: 454–467, 2008.

[29] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation revisited: Models and optimization. In *ECCV*, pages II: 465–479, 2010.

[30] J. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.