Published in final edited form as:
Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2011 ; : 1881–1888. doi:10.1109/CVPR.2011.5995420.

Scale Invariant cosegmentation for image groups*

Lopamudra Mukherjee,
Math. & Computer Science, Univ. of Wisconsin–Whitewater

Vikas Singh, and
Biostatistics & Med. Inform., Univ. of Wisconsin–Madison

Jiming Peng
Industrial & Systems Eng., Univ. of Illinois Urbana Champaign

Lopamudra Mukherjee: mukherjl@uww.edu; Vikas Singh: vsingh@biostat.wisc.edu; Jiming Peng: pengj@illinois.edu

Abstract

Our primary interest is in generalizing the problem of Cosegmentation to a large group of images, that is, concurrent segmentation of common foreground region(s) from multiple images. We further wish for our algorithm to offer scale invariance (foregrounds may have arbitrary sizes in different images) and the running time to increase (no more than) near linearly in the number of images in the set. What makes this setting particularly challenging is that even if we ignore the scale invariance desiderata, the Cosegmentation problem, as formalized in many recent papers (except [1]), is already hard to solve optimally in the two image case. A straightforward extension of such models to multiple images leads to loose relaxations; and unless we impose a distributional assumption on the appearance model, existing mechanisms for image-pair-wise measurement of foreground appearance variations lead to significantly large problem sizes (even for moderate number of images). This paper presents a surprisingly easy to implement algorithm which performs well, and satisfies all requirements listed above (scale invariance, low computational requirements, and viability for the multiple image setting). We present qualitative and technical analysis of the properties of this framework.

1. Introduction

Segmentation of an image into its constituent components (or regions) is a fundamental challenge in early vision. One typically formalizes this task as the minimization of an energy function on a graph, a strategy which goes back to papers from the 1970s [2]. In the last decade, the development of efficient graph partitioning based solutions [3, 4, 5] has led to powerful *global* methods that are robust and work well in practice. The body of work is now mature, both from a theoretical as well as a practical point of view. A majority of this literature, one must note, is focused on the segmentation of a *single* image (e.g., into foreground and background regions). In this setting, segmentation (like clustering) is an ill-posed problem. This difficulty is resolved in part by introducing an inductive bias – either via guidance from the user in the form of foreground or background seeds or by asking that the size of the two components be balanced (i.e., normalized). The goal then is to optimize the segmentation function regularized by this additional bias.

*This work is supported by UW-ICTR, UW-ADRC, and grants NIH R21-AG034315, AFOSR FA9550-09-1-0098, NSF DMS 09-15240.

Rother et al. [6] observed that the inherent ambiguity in image segmentation above can be partly mitigated if one has access to *two* images of the same object (even with different backgrounds). The so-called "Cosegmentation" problem was formalized as a simultaneous segmentation of the image pair with an additional requirement of consistency between the corresponding histograms of *only* the foreground pixels. The resultant function, in its original form, turned out to be difficult to optimize efficiently. But since then, a great deal of interest by various authors has provided progressively better algorithms [7, 1, 8, 9, 10, 11]. While there are still certain unresolved questions, efficient methods for image *pair* cosegmentation [1, 8] *are* available. Expectedly, the interest now is in generalizations of the problem – extending the formalization *beyond the two-image setting*.

Making cosegmentation viable for a *group* of images (> 2) has multiple immediate applications beyond providing a better regularization for the segmentation task. Simultaneous cutout (or identification) object of interest in an image group enables efficient editing of all occurrences in one step [11]. The authors in [8] presented experiments on building a summary collage of foregrounds from a group of images using cosegmentation, where as [1] applied the idea to identifying pathologies in brain images. Very recently, [12] showed how to leverage cosegmentation (with user intervention) to create 3D models of individuals from a few casually taken pictures – with an eye on implanting individuals in virtual environments such as video games. Preliminary versions of some of these initiatives have also been translated into end-user applications (http://chenlab.ece.cornell.edu/projects/iScribble).

1.1. Background and Related Work

We briefly review the recent set of results to place our present work in context. Cosegmentation [6] performs simultaneous segmentation of *two* images to extract the *same* foreground object. Consistency between the two foreground regions is enforced by asking that the distribution induced by the foreground pixels over a dictionary of pre-specified appearance features (such as color or textures) be similar. Let $\mathbf{x}$ denote the decision variable providing the desired segmentation. The cosegmentation energy is expressed as

$$\sum_{u=1}^{2}\left(\underbrace{\sum_{p} w_p^{(u)} x_p + \sum_{(p,q)} w_{(p,q)}^{(u)} |x_p^{(u)} - x_q^{(u)}|}_{\text{MRF terms}}\right) + \underbrace{E(h(\mathbf{x}^{(1)}), h(\mathbf{x}^{(2)}))}_{\text{global terms}}, \tag{1}$$

where the first two terms are the standard Markov Random Field (MRF) energy terms for the two images: w_p (and $w_{(p,q)}$) give the unary and pairwise terms. The functional $E(\cdot, \cdot)$ is a global term to penalize the difference between the two foreground histograms, $h(\mathbf{x}^{(1)})$ and $h(\mathbf{x}^{(2)})$ (both functions of the respective segmentation in the images), forcing the algorithm to extract mutually consistent regions.

The problem in (1) is difficult to solve efficiently for most $E(\cdot, \cdot)$ of interest – for instance, the ℓ_1-norm in [6], sum of squared differences in [7], or the dual-decomposition based method recently described in [11]; the sole exception is the submodular function optimized in [1] via a network flow procedure. (this formulation substitutes the penalty with a reward function, see [11]). Most of these papers were focused on *two* image cosegmentation but may in principle be extended to multiple images – however, even if we ignore necessary algorithmic modifications, a direct extension will involve at least a quadratic increase in the

number of additional global terms for each image pair. Further, because these terms are hard to optimize, the relaxations suffer significantly. To address these difficulties, Batra et al. [8] recently attempted to directly study the multi-image cosegmentation problem. The key idea was to avoid the main optimization related challenges by (a) making the process interactive, and (b) separating the segmentation energy of each image, tied together by a common *parametric* appearance model for all images. While leveraging user guidance (if available) is useful, the common appearance model in [8] is shared by *both* the foreground *and* background regions – a non-negligible limitation if the background varies between images in the group. Some recent works [10, 9] have presented results by restating the cosegmentation objective in terms of finding common patterns [9] or as a discriminative clustering problem [10]. The algorithm in [10] is interesting and allows incorporating color consistency and spatial constraints but requires specialized strategies to keep the problem size manageable for more than two images. For example, the kernel matrix is defined for *all* pixel pairs from *all* images –and may be an issue when more than a few images are available (for more details see [10], pp. 4). Our goal here is to study the multi-image cosegmentation problem with a focus on resolving the limitations outlined above. Specifically, the **key contributions** of this paper are: **(a)** We present a generalization of image Cosegmentation to the multi-image setting; **(b)** The model is scale invariant; further the run-time increases only linearly with the number of images to be segmented; **(c)** We provide an analysis of its theoretical and empirical properties, comparing it to the existing literature on this problem.

2. Preliminaries

We assume that a dictionary of features (based on pixel intensities, SIFT features, textures) is available for a set of l images, $\{\mathcal{I}^{(1)}, \ldots \mathcal{I}^{(l)}\}$ where each image consists of n pixels. Features in this dictionary provide $\{1, \cdots, K\}$ equivalence classes – pixels that fall in the same class are assumed to be perceptually similar. This process need not be perfect, only good enough to provide nominal guidance to the segmentation engine (methods similar to [8, 1] will suffice). As in [1], we use these equivalence classes as "bins" of a histogram. Define a matrix operator, $\mathbf{H}^{(u)}$ of size $K \times n$,

$$H^{(u)}(b, j) = \begin{cases} 1 & \text{if } I^{(u)}(j) \in \text{class } b; \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where u, j, and b index images, pixels, and histogram bins respectively. Summing over the rows of $\mathbf{H}^{(u)}$ provides the complete histogram for each image $\mathcal{I}^{(u)}$, on the pre-specified dictionary. Recall from (1) that $\mathbf{x}^{(u)}$ is the decision variable for segmentation (1 for foreground, 0 for background). One can express the corresponding foreground histogram(s), $\mathcal{F}^{(u)}$ of size $(K \times 1)$, obtained *after* segmentation:

$$\mathcal{F}^{(u)}(b) = \sum_{j=1}^{n} H^{(u)}(b, j)\mathbf{x}^{(u)}(j) \quad \forall b \in \{1, \cdots, K\}, \tag{3}$$

That is, $\mathcal{F}^{(u)} = \mathbf{H}^{(u)}\mathbf{x}^{(u)}$. With this notation in place, let us return to the issue of how existing methods enforce consistency among the two foreground histograms, concurrently with the segmentation of the two images. The global bias, $E(\mathcal{F}^{(1)}, \mathcal{F}^{(2)})$ roughly takes the form $|\mathcal{F}^{(1)} - \mathcal{F}^{(2)}|_1$ in [6], $(\mathcal{F}^{(1)} - \mathcal{F}^{(2)})^2$ in [7], and $-\langle \mathcal{F}^{(1)}, \mathcal{F}^{(2)} \rangle$ in [1], as summarized in [11]. Note that it is *this* term which makes the otherwise submodular energy more challenging to optimize. Apart from non-submodularity, the viability of such cost functions is problematic in the

context of two specific issues: **(1)** (**Scale**): Even for the basic cosegmentation setting with only two images, taking into account scale variations in the foreground will necessarily involve a grid search over $\alpha > 0$ to evaluate $E(\mathcal{F}^{(1)}, \alpha\,\mathcal{F}^{(2)})$, which is already undesirable[1]. In the multiple image setting, a search for the correct α soon becomes infeasible, even for a small number of images. **(2)** (**Multiple images**): Since $E(\cdot, \cdot)$ is introduced for *each* image-pair, the number of these terms in the objective function grow quickly with additional images. $E(\cdot, \cdot)$ is hard to optimize; in the presence of many such terms, lower bounds offered by the relaxations must deteriorate sharply, and lie farther away from the true optimum.

3. Main Ideas

Our overall goal clearly is to make the vectors $\{\mathcal{F}^{(1)}, \cdots, \mathcal{F}^{(1)}\}$ 'similar', while performing simultaneous segmentation of images $\{\mathcal{I}^{(1)}, \ldots \mathcal{I}^{(l)}\}$. The main observation is to see that 'similarity', especially in terms of how it is typically measured is a stronger than necessary requirement. In fact, it is sufficient in this setting to ask that the vectors have low conditional entropy: for images u and u', knowledge of $\mathcal{F}^{(u)}$ allows one to express $\mathcal{F}^{(u')}$ (modulo scale) with probability approaching one. As described next, this property can be formalized using the simple idea of linear dependence. From (2), we derive a vector $\hat{\mathbf{H}}^{(u)}$ of size $(K \times 1)$ where $\widehat{H}^{(u)}(b) = \sum_{j=1}^{n} H^{(u)}(b, j)$. We stack all such $\hat{\mathbf{H}}^{(u)}$'s as columns to get a matrix $\hat{\mathbf{H}}$ of size $(K \times l)$, where $\hat{H}(b, u)$ is the number of pixels in bin b from image u. The desired segmentation, $\mathbf{x} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(l)}\}$, decomposes $\hat{\mathbf{H}} = \mathcal{F} + \mathcal{B}$ as a sum of the two matrices, $\mathcal{F} = (\mathcal{F}^{(1)} \cdots \mathcal{F}^{(l)})$ and $\mathcal{B} = (\mathcal{B}^{(1)} \cdots \mathcal{B}^{(l)})$, where $\mathcal{F}$ (and $\mathcal{B}$) denote the foreground (and background) histograms respectively. In Cosegmentation, we want $\mathcal{F}^{(u)}$ and $\mathcal{F}^{(u')}$ for $u \neq u'$ to be the same (or a constant multiple of the other if the foregrounds are of different sizes). To ensure this, rather than explicitly penalize the gross difference between $\mathcal{F}^{(u)}$ and $\mathcal{F}^{(u')}$, we can ask that the columns of $\mathcal{F}$ have low entropy and be linearly *dependent*. An important consequence of this property is that it is *completely immune* to scale variations, exactly as desired in cosegmentation. An effective means for making this idea operational is to ensure that the *rank of* $\mathcal{F}$ *is one*. In cases where a precise rank one $\mathcal{F}$ cannot be found, a "slack" in the form of a small (sparse) residual, $\mathcal{L}$ may be permitted. We may now define the following model

$$\begin{aligned} \min_{\mathbf{x},\mathcal{P},\mathcal{L}} \quad & \sum_u \mathbf{w}_p^{(u)} \mathbf{x}^{(u)} + \sum_{u,(p\sim q)} \mathbf{w}_{(p\sim q)}^{(u)} \left| x_p^{(u)} - x_q^{(u)} \right| + \|\mathcal{L}\|_2^2 \\ \text{s.t.} \quad & \mathbf{H}^{(u)} \mathbf{x}^{(u)} = \mathcal{F}^{(u)} \quad \forall u \\ & \mathcal{P} + \mathcal{L} = \mathcal{F}, \quad \text{rank}(\mathcal{P}) = 1, \end{aligned} \tag{4}$$

where $\mathbf{w}_p$ (and $\mathbf{w}_{(p\sim q)}$) are the vector representations of the unary (and pairwise smoothness) terms in (1), and $\mathbf{H}^{(u)}$ is of size $(K \times n)$. The objective penalizes the residual to keep the variation (from a rank one matrix) small. Note that the constraint $\hat{\mathbf{H}} = \mathcal{F} + \mathcal{B}$ may be included, but is redundant.

4. Cosegmentation for image groups

Using $(\mathcal{P} \approx \mathcal{F}) \leq \hat{\mathbf{H}}$, we may consider a model which penalizes the slack version of the constraint, $\mathcal{P} + \mathcal{L} = \mathcal{F}$,

[1][11] reported that using an extension of the Boykov-Jolly model [13] provides some robustness to small scale differences in experiments.

$$\min \quad \sum_u \left(\mathbf{w}_p^{(u)} \mathbf{x}^{(u)} + \mathbf{w}_{(p\sim q)}^{(u)} \left| x_p^{(u)} - x_q^{(u)} \right| \right) + \lambda \left| \underbrace{\begin{bmatrix} \vdots & & \vdots \\ \mathbf{H}^{(1)}\mathbf{x}^{(1)} & \cdots & \mathbf{H}^{(l)}\mathbf{x}^{(l)} \\ \vdots & & \vdots \end{bmatrix}}_{\mathcal{F}} - \mathcal{P} \right|_2^2 \quad \text{rank}(\mathcal{P}) = 1. \tag{5}$$

We present the following simple scheme for problem (5).

A brief explanation here is useful. The constant λ balances the influence of the global terms and the segmentation terms. In (**Step 2**), for a fixed $\mathcal{P}$, we solve the problem,

$$\min_{\mathbf{x}} \sum_u \left(\mathbf{w}_p^{(u)} \mathbf{x}^{(u)} + \mathbf{w}_{(p\sim q)}^{(u)} \left| x_p^{(u)} - x_q^{(u)} \right| + \lambda \left| \mathbf{H}^{(u)} \mathbf{x}^{(u)} - \mathcal{P}^{(u)} \right|_2^2 \right), \tag{6}$$

where $\mathcal{P}^{(u)}$ is the u-th column of $\mathcal{P}$. This objective function, after some modifications can be expressed as a Quadratic Pseudoboolean function [14, 15]. This ensures that the solution is partially optimal (half integral), that is, every entry of $\mathbf{x}^*$ is $\{0, \frac{1}{2}, 1\}$. Another advantage is that for $u \neq u'$, we see that $\mathbf{x}^{(u)}$ has no interaction with $\mathbf{x}^{(u')}$. Therefore, the optimization can be performed independently for each u. Now, in (**Step 3**), for a fixed $\mathbf{x}_{[k]}$ we are given the matrix $\mathcal{F}_{[k]}$ and want to find its closest rank one matrix $\mathcal{P}_{[k+1]}$. This can be obtained by the singular value decomposition of $\mathcal{F}_{[k]}$. In addition, we may also add a constraint $\mathcal{P} \leq \hat{\mathbf{H}}$ to upper bound $\mathcal{P}$ entry-wise[2], but in most iterations this is not needed.

Lemma 1 (Monotonicity)—The above algorithm reduces the objective value of (5) at each iteration.

Proof—Denote the objective in (5) as $\mathcal{E} = \mathcal{E}^M + \mathcal{E}^G$, where $\mathcal{E}^M$ denotes the MRF terms in (5), and $\mathcal{E}^G = \lambda \left\| \mathcal{F} - \mathcal{P} \right\|_2^2$ gives the global histogram terms. The algorithm begins with a fixed $\mathcal{P}_{[1]}$ and then finds a configuration $\mathbf{x}_{[1]}$ which minimizes the function for the given $\mathcal{P}_{[1]}$. Let the energy be denoted by $\mathcal{E}_{[1]} = \mathcal{E}^M(\mathbf{x}_{[1]}) + \lambda \left\| \mathcal{F}_{[1]} - \mathcal{P}_{[1]} \right\|_2^2$, where the subscript $[k]$ gives the iteration step. Note that $\mathcal{P}_{[1]}$ is a rank one matrix but it is not the closest rank one approximation of $\mathcal{F}_{[1]}$. Therefore, the objective can be further improved by replacing $\mathcal{P}_{[1]}$ by $\mathcal{P}_{[2]}$, where $\mathcal{P}_{[2]}$ *is* the rank one approximation of $\mathcal{F}_{[1]}$ (e.g., computed as in Step 3 above). Let the new objective value be $\widehat{\mathcal{E}_{[1]}} = \mathcal{E}^M(\mathbf{x}_{[1]}) + \lambda \left\| \mathcal{F}_{[1]} - \mathcal{P}_{[2]} \right\|_2^2$. Clearly, $\mathcal{E}_{[1]} \geq \widehat{\mathcal{E}_{[1]}}$. Now, keeping $\mathcal{P}_{[2]}$ fixed we solve and obtain a new configuration $\mathbf{x}_{[2]}$. The objective value now is $\mathcal{E}_{[2]} = \mathcal{E}^M(\mathbf{x}_{[2]}) + \lambda \left\| \mathcal{F}_{[2]} - \mathcal{P}_{[2]} \right\|_2^2$. Since $\mathbf{x}_{[2]}$ is the optimal configuration for $\mathcal{P}_{[2]}$, we have $\mathcal{E}_{[1]} \geq \mathcal{E}_{[2]}$. This argument applies to any two consecutive configurations, $\mathbf{x}_{[k]}$ and $\mathbf{x}_{[k+1]}$, and so the objective value must be monotonically non-increasing at each iteration.

[2]A small technicality arises when $\mathcal{P}_{[k+1]} \leq \hat{\mathbf{H}}$ is not satisfied entry-wise. In this case, the singular vectors can be adjusted locally to satisfy the constraints, in which case we are using the best rank one approximation which also satisfies the entry-wise inequalities.

Theorem 1 (Stationary Point)—The above algorithm will converge after a finite number of steps to a feasible solution.

Proof—The function $\mathcal{E}(\mathbf{x})=\mathcal{E}^M(\mathbf{x})+\lambda\left|\left|\mathcal{F}-\mathcal{P}\right|\right|_2^2$ has half-integral solutions if $\mathcal{P}$ is fixed (see Step 2 above). This means that the solution values of any configuration $\mathbf{x}_{[k]}$ can only be drawn from $\{0, \frac{1}{2}, 1\}$, and the number of configurations for the problem is finite. Because the objective value from Lemma 1 is monotonically non-increasing, and only a finite set of configurations can be visited (and each $\mathcal{P}$ is rank one), the method must converge to a feasible stationary point.[3]

4.1. Properties of the Hessian

An alternative way of writing the objective in (5) is replacing the constraint rank($\mathcal{P}$) = 1 as $\mathcal{P}=\mathbf{u}\mathbf{v}^T$ where **u** (this is a different variable from the non-bold face *u* used in §3) (and **v**) are vectors of size $K \times 1$ (and $l \times 1$) respectively. Then we can rewrite our objective function as

$$f(\mathbf{x},\mathbf{u},\mathbf{v})=\mathcal{E}^M(\mathbf{x})+\lambda|\mathcal{F}-\mathbf{u}\mathbf{v}^T|_2^2. \tag{7}$$

While the method in Fig. 1 is sufficient in practice, because of the implicit rank one constraint, the problem remains non-convex and difficult to solve to global optimality. The standard practice in such situations is to check whether the stationary point obtained from the iterative procedure satisfies second order conditions, which means a local minimum has been achieved [16]. In other words, we need to check whether the Hessian (say, *D*) of the objective function at the stationary point is positive semidefinite. If $D \not\succeq 0$, we must find a new search direction to further reduce the objective value. Next, we explore the conditions under which the Hessian is positive semidefinite. For notational convenience, we omit λ. Using (7), we express our objective as

$$\begin{aligned} f(\mathbf{x},\mathbf{u},\mathbf{v}) &= \mathcal{E}^M(\mathbf{x})+\mathrm{tr}(\mathcal{F}^T\mathcal{F}-2\mathcal{F}^T\mathbf{u}\mathbf{v}^T+\mathbf{v}\mathbf{u}^T\mathbf{u}\mathbf{v}^T) \\ &= \mathcal{E}^M(\mathbf{x})+\sum_u \mathbf{x}^{(u)^T}\mathbf{H}^{(u)^T}\mathbf{H}^{(u)}\mathbf{x}^{(u)}-2\mathbf{u}^T\mathcal{F}\mathbf{v}+\mathbf{u}^T\mathbf{u}\mathbf{v}^T\mathbf{v}) \end{aligned}$$

From the above relation, we derive the symmetric Hessian matrix of the objective function in (8) as follows focusing only on the global terms (MRF terms are convex)

$$\begin{bmatrix} \frac{\partial^2 f}{\partial \mathbf{x}^2} & \frac{\partial^2 f}{\partial \mathbf{x}\partial \mathbf{u}} & \frac{\partial^2 f}{\partial \mathbf{x}\partial \mathbf{v}} \\ \frac{\partial^2 f}{\partial \mathbf{u}\partial \mathbf{x}} & \frac{\partial^2 f}{\partial \mathbf{u}^2} & \frac{\partial^2 f}{\partial \mathbf{u}\partial \mathbf{v}} \\ \frac{\partial^2 f}{\partial \mathbf{v}\partial \mathbf{x}} & \frac{\partial^2 f}{\partial \mathbf{v}\partial \mathbf{u}} & \frac{\partial^2 f}{\partial \mathbf{v}^2} \end{bmatrix} = 2\begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{pmatrix} = 2D, \tag{8}$$

where

[3]If the problem has multiple optimal solutions, to rule out oscillations, one may use a deterministic procedure to break ties when selecting the most violated constraints (e.g., in the simplex procedure).

$$D_{11}=\mathrm{Diag}(\mathbf{H}^{(1)^T}\mathbf{H}^{(1)},\mathbf{H}^{(2)^T}\mathbf{H}^{(2)},\cdots,\mathbf{H}^{(l)^T}\mathbf{H}^{(l)}),$$
$$D_{22}=\mathbf{v}^T\mathbf{v}\mathbf{I}_K,\quad D_{21}=D_{12}^T=-(\mathbf{H}^{(1)}v_1,\mathbf{H}^{(2)}v_2\cdots,\mathbf{H}^{(l)}v_l),$$
$$D_{32}=D_{23}^T=-\mathcal{F}^T+2\mathbf{v}\mathbf{u}^T,\ D_{33}=\mathbf{u}^T\mathbf{u}\mathbf{I}_l$$
$$D_{31}=D_{13}^T=-\mathrm{Diag}(\mathbf{u}^T\mathbf{H}^{(1)},\mathbf{u}^T\mathbf{H}^{(2)},\cdots,\mathbf{u}^T\mathbf{H}^{(l)}).$$

Here, $\mathbf{I}_p$ gives a $p \times p$ identity matrix, and v_i is the i-th entry of $\mathbf{v}$ (similarly for $\mathbf{u}$). Since each $\mathbf{H}^{(u)}$ is of size $(K \times n)$, D_{11} is of size $(nl \times nl)$. Let us define

$$Q=\begin{pmatrix}\mathbf{I}_{nl} & 0 & -\frac{D_{13}}{\mathbf{u}^T\mathbf{u}}\\ 0 & \mathbf{I}_K & -\frac{D_{23}}{\mathbf{u}^T\mathbf{u}}\\ 0 & 0 & \mathbf{I}_l\end{pmatrix}. \tag{9}$$

For blocks $\tilde{D}_{11}$, $\tilde{D}_{12}$, $\tilde{D}_{21}$, and $\tilde{D}_{22}$, it can be verified that

$$QDQ^T-\begin{pmatrix}\tilde{D}_{11} & \tilde{D}_{12} & 0\\ \tilde{D}_{21} & \tilde{D}_{22} & 0\\ 0 & 0 & D_{33}\end{pmatrix}. \tag{10}$$

Because Q is nonsingular, the above relation implies

Lemma 2—The Hessian D satisfies $D \succcurlyeq 0$ if and only if

$$\tilde{D}=\begin{pmatrix}\tilde{D}_{11} & \tilde{D}_{12}\\ \tilde{D}_{21} & \tilde{D}_{22}\end{pmatrix}\succcurlyeq 0. \tag{11}$$

To explore the properties of the matrix $\tilde{D}$, we first show the following result

Lemma 3—The blocks $\tilde{D}_{11}$ and $\tilde{D}_{22}$ in $\tilde{D}$ satisfy $\tilde{D}_{11} \succcurlyeq 0$ and $\tilde{D}_{22} \succcurlyeq 0$.

Proof—Consider the singular value decomposition of $\mathcal{F}$,

$$\mathcal{F}=\sum_{i=1}^{\min(K,l)}\sigma_i\bar{u}^i(\bar{v}^i)^T,\quad \left\|\bar{u}^i\right\|=\left\|\bar{v}^i\right\|=1,\quad \forall i;$$

where $\bar{u}^i$, and $\bar{v}^i$ are the left-hand and right-hand side singular vectors corresponding to the singular values, σ_i of $\mathcal{F}$, and σ_1 is its largest singular value. It follows that

$$(\bar{u}^i)^T\bar{u}^j=0,\quad (\bar{v}^i)^T\bar{v}^j=0,\quad \forall i\neq j;$$
$$\mathbf{u}=\sqrt{\sigma_1}\bar{u}^1,\quad \mathbf{v}=\sqrt{\sigma_1}\bar{v}^1.$$

This implies that $(-\mathcal{F}+2\mathbf{u}\mathbf{v}^T)=\mathbf{u}\mathbf{v}^T - \sum_{i=2}^{\min(K,l)} \sigma_i \overline{u}^i (\overline{v}^i)^T$. Therefore,

$$\begin{aligned}\tilde{D}_{22}=\sigma_1\mathbf{I}_k - \tfrac{1}{\sigma_1} D_{23}D_{32} &= \sigma_1\mathbf{I}_K - \tfrac{1}{\sigma_1}\sum_{i=1}^{\min(K,l)} \sigma_i^2 \overline{u}^i (\overline{u}^i)^T \\ &\succcurlyeq \sum_{i=2}^{\min(K,l)} \tfrac{\sigma_1^2-\sigma_i^2}{\sigma_1} \overline{u}^i (\overline{u}^i)^T \succcurlyeq 0.\end{aligned} \tag{12}$$

Similarly,

$$\begin{aligned}\tilde{D}_{11}=\mathrm{Diag}(\mathbf{H}^{(1)^T}(\mathbf{I}_K - \tfrac{\mathbf{u}\mathbf{u}^T}{\sigma_1})\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(l)^T}(\mathbf{I}_K - \tfrac{\mathbf{u}\mathbf{u}^T}{\sigma_1})\mathbf{H}^{(l)}) \\ \succcurlyeq 0,\end{aligned} \tag{13}$$

providing the desired result.

We now obtain an identity for $\tilde{D}_{12}$, which will be helpful shortly. Let us denote the i-th row of a matrix M as $[\![M]\!]$ $(i\cdot)$. From (11), we get $\tilde{D}_{12}=D_{12} - \frac{1}{\sigma_1} D_{13}D_{32}$. This yields

$$\begin{aligned}\tilde{D}_{12}(i\cdot) &= -\mathbf{H}^{(i)^T} v_i + \tfrac{1}{\sigma_1}\mathbf{H}^{(i)^T}\mathbf{u} - \mathcal{F}^T + 2\mathbf{v}\mathbf{u}^T(i\cdot) \\ &= -\mathbf{H}^{(i)^T} v_i + \tfrac{1}{\sigma_1}\mathbf{H}^{(i)^T}\mathbf{u}\mathbf{v}\mathbf{u}^T - \sum_{j=2}^{\min(K,l)} \sigma_j \overline{v}^j (\overline{u}^j)^T(i\cdot) \\ &= -v_i\mathbf{H}^{(i)^T}(\mathbf{I}_K - \tfrac{\mathbf{u}\mathbf{u}^T}{\sigma_1}) - \tfrac{1}{\sigma_1}\mathbf{H}^{(i)^T}\sum_{j=2}^{\min(K,l)} \sigma_j \overline{v}_i^j \mathbf{u}(\overline{u}^j)^T\end{aligned} \tag{14}$$

Define $\mathbf{S} = \mathrm{Diag}(\mathbf{H}^{(1)^T}, \cdots, \mathbf{H}^{(l)^T}, \mathbf{I}_K)$; we can write

$$\begin{aligned}&\tilde{D}=\mathbf{S}\widehat{D}\mathbf{S}^T, \text{ where } \widehat{D}=\begin{pmatrix}\widehat{D}_{11} & \widehat{D}_{12}\\ \widehat{D}_{21} & \widehat{D}_{22}\end{pmatrix} \text{ and} \\ &\widehat{D}_{11}=\mathrm{Diag}(\mathbf{I}_K - \tfrac{\mathbf{u}\mathbf{u}^T}{\sigma_1}, \cdots, \mathbf{I}_K - \tfrac{\mathbf{u}\mathbf{u}^T}{\sigma_1}), \widehat{D}_{22}=\tilde{D}_{22}; \\ &\widehat{D}_{12}(i\cdot)= \underbrace{-v_i(\mathbf{I}_K - \frac{\mathbf{u}\mathbf{u}^T}{\sigma_1})}_{\widehat{D}_{12}^{i(1)}} \underbrace{- \frac{1}{\sigma_1}\sum_{j=2}^{\min(K,l)} \sigma_j \overline{v}_i^j \mathbf{u}(\overline{u}^j)^T}_{\widehat{D}_{12}^{i(2)}}.\end{aligned} \tag{15}$$

If $\mathbf{S}$ is full rank, then we can conclude

Property 1

$$D \succcurlyeq 0 \textit{ iff } \widehat{D}_{22} - \sum_{i=1}^{l} \widehat{D}_{12}(i\cdot)^T \widehat{D}_{12}(i\cdot) \succcurlyeq 0.$$

Theorem 2—The Hessian of the objective function is positive semidefinite if $\mathcal{F}$ is a rank one matrix.

Proof—Recall that

$$\widehat{D}_{12}^{i(1)}\widehat{D}_{12}^{i(2)}=0, \tag{16}$$

it follows

$$\sum_{i=1}^{l}\widehat{D}_{12}^{i}{}^{T}\widehat{D}_{12}^{i}=\sum_{i=1}^{l}v_i^2(\mathbf{I}_K-\frac{\mathbf{u}\mathbf{u}^T}{\sigma_1})+\sum_{i=2}^{l}\widehat{D}_{12}^{i(2)^T}\widehat{D}_{12}^{i(2)} \tag{17}$$

$$\geqslant \sigma_1(\mathbf{I}_k-\frac{\mathbf{u}\mathbf{u}^T}{\sigma_1}). \tag{18}$$

The above relation implies

$$\widehat{D}_{22}-\sum_{i=1}^{l}\widehat{D}_{12}^{i}{}^{T}\widehat{D}_{12}^{i}\leqslant-\sum_{i=2}^{\min(K,l)}\frac{\sigma_i^2}{\sigma_1}\overline{u}^i(\overline{u}^i)^T\leqslant 0; \tag{19}$$

equality holds when $\sigma_j = 0$ for all $j = 2, \cdots, l$.

Remark—We see from Thm. 2 that if $\mathcal{F}$ is not a rank-1 matrix, then the Hessian D is not positive semidefinite. In practice, we found that the negative eigen-values were extremely small and so the hessian was *nearly* positive semidefinite. Indeed, (19) suggests that the variation from a rank one matrix becomes progressively small as the first eigen value dominates the others. In fact, even if D has negative eigen-values, they are guaranteed to lie in the lower order block (lowest $l - 1$). This can be shown by a simpler perturbation theoretic analysis (see longer version of the paper).

5. Experiments

Our experimental evaluations described next were designed to assess the algorithm's performance vis-à-vis three other Cosegmentation methods [7, 10, 7]. For the base case, in §5.1 we evaluate consistency of our segmentations (for an image *pair*) with other methods designed specifically for two images. Then, §5.1 shows our results for image *groups*, relative to those obtained from [10] (a recent method which also works for multiple images). Later, results in §5.2 show robustness to foreground scale differences, and §5.4 sheds additional light on various experimental properties of the model. Finally, §5.5 assesses the algorithm's running time requirements, to conclude our empirical analysis. **Datasets:** We conducted experiments on a set of 25 groups of images collected from various sources (over 100 images in all). These include image pairs used in previous work [6, 1] (both publicly available), as well as larger image sets used in [8, 9]. In addition, similar to [10], our image groups include images from Weizman horses[4], MSRC object categories[5], and Oxford

[4]www.msri.org/people/members/eranb/
[5]research.microsoft.com/en-us/projects/objectclassrecognition/

flowers dataset[6]. Since several categories included high variability images with similar background as well as foreground, we used a few seed points to facilitate the segmentation process. Consistent with existing works [1, 10], histograms were generated using a combination of color, texture features, and SIFT. The number of bins for each color channel was between 10–20.

5.1. Base case: two image cosegmentation

We first evaluate how our new objective function proposed (for multiple images) compares to more traditional formulations of Cosegmentation [7, 1] that are specifically designed to segment two images at a time. We evaluate all three methods on a collection of 10 pairs of images, examples are shown in Fig. 2. Qualitatively, the results show that our method is competitive to these methods that have been shown to be partially optimal [7] and optimal [1] for their respective objectives. For the first image pair, our results are the best where as they are only slightly worse than [7, 1] on the other images. Measuring accuracy by the proportion of incorrectly classified pixels to the total number of pixels, for the examples in Fig. 2, we obtain the following error estimates:

error	woman	stone	can
Mukherjee [7]	2.14%	1.5%	**2.3**%
Hochbaum [1]	4.0%	**1.2**%	7.1%
Our method	**1.5**%	4.6%	4.7%

The trend is mixed, which is representative of the remaining images. Overall, the average of all three algorithms was under 5%, and our method was within 1.5% of either method.

5.2. Scale Invariance

Our algorithm, by design, offers scale-invariant Cosegmentation. To see how this plays out in practice, we ran comparisons on image sets where the foreground regions were significantly different in size. As a baseline, we also evaluated the algorithms from [1], and to the recently proposed method in [10] (using an implementation provided by the authors). Unlike our model, these and other methods do not explicitly account for scale. Fig. 3 shows two such examples where the proposed method is able to identify substantial variations in foreground scale without difficulty. In comparison, the algorithms from [10, 1] show oversegmentation, especially in the *second* image (which has a smaller foreground region).

5.3. Cosegmentation for image groups

Next, we compared our method in the multi image cosegmentation setting with the algorithm from [10]. We did not perform comparisons with [8] because it assumes the same appearance model for both the background and foreground; this creates difficulties for many images considered here (unless sufficient user interaction is incorporated). Comparisons were carried out on 15 groups of images, with varied number of images in each group. Some representative samples are shown in Fig. 4. In general, we see that while in some image groups, the performance of both methods are comparable (e.g., Oxford flowers and dog), in some other images like banana (which have a shared background in some images), the

[6] www.robots.ox.ac.uk/~vgg/data/flowers/17/

algorithm in [10] oversegments. Ignoring such images (where the difference in performance is higher), the accuracy of our (and [10]) was 4.6 ± 2.1% (and 6.9 ± 1.8%) respectively.

5.4. Other empirical properties of the model

Theorem 2 shows that the ratio of the first and the second eigen value will modulate the behavior of our algorithm in practice. In our experiments, we observe that this ratio starts out around 0.9 and decreases in progressive steps until it stabilizes. The decay of this ratio is faster for fewer images; this is expected since the rank of the initial histogram matrix directly depends on the number of input images. Also, our analysis of the many (problem instance specific) Hessians shows that the negative eigen values (when they exist) are very small. As a result, if one decides to make use of a second order method, modifications to the Hessian will be minimal (if any). We did not find such a procedure necessary.

5.5. Running time

The algorithm takes fewer than 10 iterations to converge in all cases tested. One iteration of solving (6) on a standard workstation (using network flow) takes 10 – 30*s* on a 128 × 128 image (the variation is based on the number of histogram bins). For two images, this is comparable (within a factor of two) to methods presented in our experiments above [1, 7]. However, a significant advantage of the method becomes apparent in the multiple image setting. Observe that with an increase in the number of images in the group, the running time increase of our model is only *additive* since we solve for $\mathbf{x}^{(u)}$ for each u individually, note that we do not need to perform the complete singular value decomposition, but only its rank k approximation. The running is time of such operations in $O(kmn)$, see [17], which is linear in n (the number of images). On the other hand, for the method in [10] the problem size increases quickly with more images, and the reported running time in [10] for 30 images was 4–9 hours (calculated under the setting where superpixels were used). Our proposed method, therefore, offers substantial improvements in running time with the additional advantage of scale invariance.

6. Conclusions

We have presented a generalization of the Cosegmentation problem to image *groups*, which is also immune to foreground scale variations (multiple objects in the foreground are permitted as long as there exists a constant scale factor relating the foregrounds for an image pair). The algorithm is easy to implement and has a small computational footprint – the run-time increases only linearly with additional images (such an increase cannot be avoided). We have also provided a technical and empirical analysis of the properties of this model, together with preliminary qualitative and quantitative evidence to demonstrate that the algorithm performs well in practice.

References

1. Hochbaum D, Singh V. An efficient algorithm for co-segmentation. Intl Conf on Comp Vis. 2009
2. Zahn CT. Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computing. 1971; 20(1):68–86.
3. Ishikawa H, Geiger D. Segmentation by grouping junctions. Comp Vision and Pattern Recog. 1998:125.
4. Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. Trans on Pattern Anal and Machine Intel. 2001; 23(11):1222–1239.
5. Shi J, Malik J. Normalized cuts and image segmentation. Trans on Pattern Analysis and Machine Intell. 2000; 22(8):888–905.

6. Rother C, Minka T, Blake A, Kolmogorov V. Cosegmentation of image pairs by histogram matching: Incorporating a global constraint into MRFs. Comp Vision and Pattern Recog. 2006
7. Mukherjee L, Singh V, Dyer C. Half-integrality based algorithms for cosegmentation of images. Comp Vision and Pattern Recog. 2009
8. Batra D, Kowdle A, Parikh D, Luo J, Chen T. iCoseg: Interactive cosegmentation with intelligent scribble guidance. Comp Vision and Patter Recog. 2010
9. Chu WS, Chen CP, Chen CS. MOMI-cosegmentation: Simultaneous segmentation of multiple objects among multiple images. Asian Conf on Comp Vision. 2010
10. Joulin A, Bach F, Ponce J. Discriminative clustering for image cosegmentation. Comp Vision and Patter Recog. 2010
11. Vicente S, Kolmogorov V, Rother C. Cosegmentation Revisited: Models and Optimization. European Conf on Comp Vision. 2010
12. Kowdle A, Batra D, Chen WC, Chen T. iModel: Interactive cosegmentation for object of interest 3d modeling. European Conf on Comp Vision Workshop. 2010
13. Boykov Y, Jolly M. Interactive graph cuts segmentation of objects in n-d images. Intl Conf on Comp Vis. 2001
14. Boros E, Hammer P. Pseudo-Boolean optimization. Disc Appl Math. 2002; 123:155–225.
15. Rother C, Kolmogorov V, Lempitsky V, Szummer M. Optimizing binary mrfs via extended roof duality. Comp Vision and Pattern Recog. 2007
16. Nocedal, J.; Wright, SJ. Numerical optimization. Springer Verlag; 1999.
17. Frieze AM, Kannan R, Vempala S. Fast monte-carlo algorithms for finding low-rank approximations. Journal of ACM. 1998; 51(6):1025–1041.

Algorithm

Step 1) (Set $k = 1$). Initialize $\mathbf{x}_{[0]}$ to $\mathbf{1}$. Select a rank one matrix $\mathcal{P}_{[k]} \leq \hat{\mathbf{H}}$.

Step 2) Solve Problem (5) for fixed $\mathcal{P}_{[k]}$, and denote the optimal solution by $\mathbf{x}_{[k]}$.

Step 3) Solve Problem (5) for fixed $\mathbf{x}_{[k]}$; denote the optimal solution by $\mathcal{P}_{[k+1]}$.

Step 4) If $\|\mathbf{x}_{[k]} - \mathbf{x}_{[k-1]}\| \leq \epsilon$, stop. Otherwise set $k = k + 1$, go back to step 2.

Figure 1.
Our proposed algorithm to optimize (5).

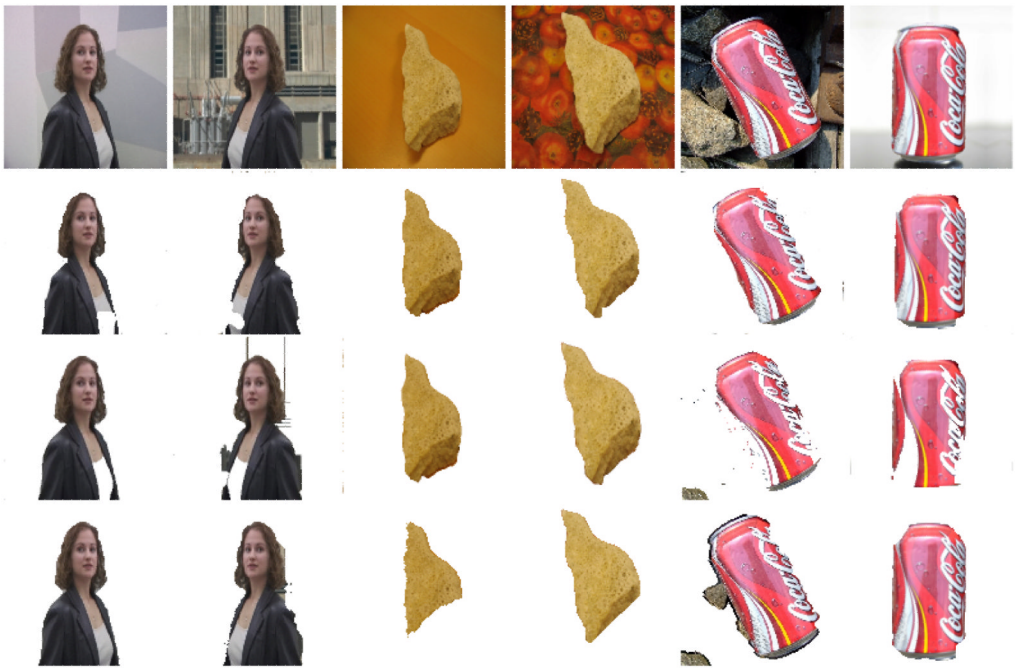

Figure 2.
Representative examples from a comparison of the proposed multi-image cosegmentation with the algorithms proposed in [7, 1]. Row 1 shows the original image pairs; Rows 2–3 gives the result from [7] and [1] respectively, and our solution is shown in Row 4.

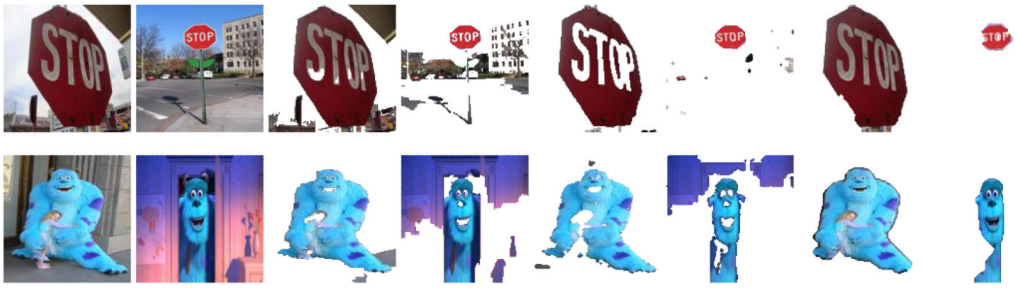

Figure 3.
Example input images with significant differences in foreground size are shown. Columns 3–4 gives results from [10]; Columns 5–6 show solutions from [1]. The last two columns present our results on these image pairs.

Figure 4.
Comparison results of the algorithm proposed in [10] (in columns 2,5) with our method (in columns 3,6). Original images provided in columns 1,4.