Cosegmentation of Image Pairs by Histogram Matching — Incorporating a Global Constraint into MRFs

Carsten Rother[1], Vladimir Kolmogorov[2], Tom Minka[1], and Andrew Blake[1]
[1] Microsoft Research Cambridge, {carrot, minka, ablake}@microsoft.com
[2] University College London, vnk@adastral.ucl.ac.uk
http://research.microsoft.com/vision/cambridge/

Abstract

We introduce the term cosegmentation which denotes the task of segmenting simultaneously the common parts of an image pair. A generative model for cosegmentation is presented. Inference in the model leads to minimizing an energy with an MRF term encoding spatial coherency and a global constraint which attempts to match the appearance histograms of the common parts. This energy has not been proposed previously and its optimization is challenging and NP-hard. For this problem a novel optimization scheme which we call trust region graph cuts *is presented. We demonstrate that this framework has the potential to improve a wide range of research: Object driven image retrieval, video tracking and segmentation, and interactive image editing. The power of the framework lies in its generality, the common part can be a rigid/non-rigid object (or scene), observed from different viewpoints or even similar objects of the same class.*

1. Introduction

This paper looks at segmentation, which is a fundamental problem in computer vision, and particularly at the simultaneous segmentation of a pair of images, an operation that we term "cosegmentation". Powerful procedures for low-level segmentation can be produced by incorporating difference measures at the level of pixels, into a global objective function [20, 3, 17]. The objective function can also incorporate a tendency to coherence of regions. Completely automatic segmentation is possible [20] but prone to error, and interactive input [3, 17] or fusion with other modalities [13], is normally needed to correct those errors. Another source of information for correcting segmentation is to supply a database of related images and segment them simultaneously [21]. Here we demonstrate that supplying just *one* additional image can be sufficient to segment both together, to higher accuracy than is achieved with either one alone. Furthermore, in contrast to [21] we do not exploit a shared shape model which has the advantage of being completely viewpoint independent.

Apart from clear applications in interactive graphics, for segmentation of images and videos, cosegmentation has implications in another important area: image similarity measures. Commonly the degree of similarity of a pair of images has been computed by comparing the global statistics of the two images. Typically the comparison is applied to the histograms of each image, constructed from features such as colour and texture [11, 7]. However, such a global, undifferentiated approach to comparison is liable to result in crude comparisons, as figures 5, 6 show. Apparently, it is essential to incorporate some form of differentiation of parts of images, so that comparison can be based on those parts of an image pair which are shared in common. In that way, a similarity between subjects can be scored highly, without unreasonable dilution by differences in backgrounds. (Conversely, similarities in the background scenes of a pair of images could be captured despite the subjects being unrelated.) One approach to such differentiation, is "integrated region matching" [11], in which images are subjected to mean-shift segmentation [5], and then a simple similarity measure records the similarity of paired regions, in a search over both segmented images. However, the choice of paired regions takes no account of object coherence, and so cannot properly take account of the distinction between subject and background. Here we address that shortcoming by *jointly* cosegmentation the image pair using a proper MRF coherence prior and a histogram matching cost, and then compare either subject or background.

A sub-problem which arises in cosegmentation is the problem of finding a coherent image region with given target histogram. This problem has been approached previously using ellipses or active contours to define coherence [6, 12, 9]. Inspired by [17], we instead define coherence via MRF priors and solve the problem with iterated graph cuts.

In order to arrive at an objective function for cosegmentation, we begin, in sec. 2, by setting out a generative model for an image pair, and then evaluating the hypothesis that

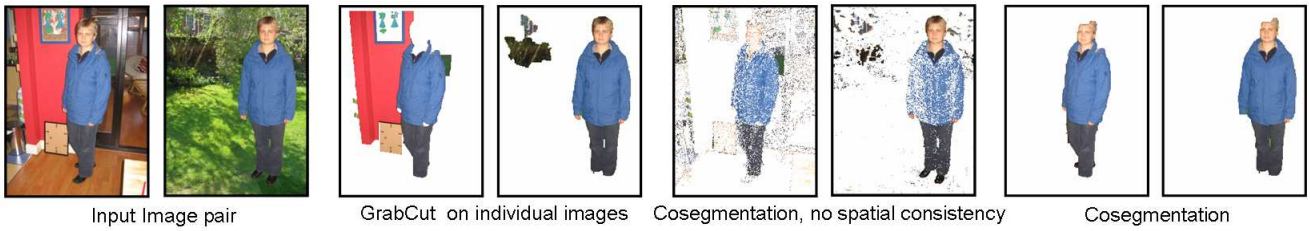

Figure 1. **Introducing cosegmentation.** Given a pair of images (a) the objective is to segment the common part in both images. (b) Shows the result of applying GrabCut [17] on the images separately, with a preference of foreground being more likely in the image center. The result is as expected since the joint foreground is not modeled. (c) Shows the result of performing cosegmentation, however, without any spatial constraints. (d) Result of our complete cosegmentation framework.

the images share common material. The recovered cosegmentation will then be that pair of regions, one from each image, under which that hypothesis is most probable. One approach to the generative model considers pixels in the backgrounds and foregrounds of each image to have been generated independently from a certain probability distribution for colour (or texture). Then, under the hypothesis, the foreground distributions are constrained to be identical. This can be shown to yield, as a likelihood for the images, a function of the well-known Jensen-Shannon divergence between foreground histograms (see [18]). However, the independence assumption is something of a drawback, as it is known that nearby pixels in an image are not generally independent [8]. If instead we choose a generative model for the foreground histograms as a whole, rather than individual pixels, we can obtain other standard divergences such as variational distance. A further Ising prior on segmentations, gated by image contrast [3], encourages smooth boundaries.

The optimisation of the objective function arising from that generative model, is something of a challenge. Graph cut algorithms are widely used for binary optimisation in Markov models [10, 3], but have not been used before where the objective function contains a histogram difference measure. It transpires that such an objective function is not "submodular" and therefore strictly not tractable. Therefore we develop, in sec. 3, a new, approximate algorithm based on graph cuts. Finally, in sec. 4, we show a series of results, demonstrating the effectiveness of the new model and algorithm in image segmentation, and in the development of image similarity measures that respect the distinction between subject and background.

2. A Generative Model for Cosegmenting Image Pairs

Due to lack of space the following derivation includes several approximations and one particular image generation model, the full derivation including an alternative image generation model can be found in [18].

Let $k \in \{1, 2\}$ range over images and $i \in \{1, \ldots, n\}$ range over pixels.

- $x_{ki} \in \{0, 1\}$ indicates whether pixel i in image k is foreground. $\mathbf{x}_k$ is shorthand for the entire labeling in image k, and $\bar{\mathbf{x}}$ is shorthand for both images.
- z_{ki} is an image measurement, e.g. colour or texture at pixel i in image k. We assume that this measurement falls into a finite number of bins. Symbol z will range over these bins. Given $\mathbf{x}_k$, z_{kf} is shorthand for all foreground pixels, and z_{kb} for all background pixels. $\mathbf{z}_k$ is shorthand for the entire image k, and $\bar{\mathbf{z}}$ is shorthand for all images.
- $\boldsymbol{\theta}_{kf}$ denotes foreground model parameters for $\mathbf{z}_k$. $\boldsymbol{\theta}_{kb}$ denotes background model parameters. $\boldsymbol{\theta}_k$ is shorthand for both $(\boldsymbol{\theta}_{kf}, \boldsymbol{\theta}_{kb})$, and $\bar{\boldsymbol{\theta}}$ is shorthand for all $\boldsymbol{\theta}_k$.

Given two images $\bar{\mathbf{z}} = (\mathbf{z}_1, \mathbf{z}_2)$, we consider two possible generative models, illustrated in Fig. 2. In both models, the segmentations and background models are independent across images. If $J = 0$ then the foreground models are independent; if $J = 1$ then the foreground models are the same. This difference shows up only in the prior for $\bar{\boldsymbol{\theta}}$. Therefore the image model given the segmentations is:

$$p(\bar{\mathbf{z}}|J, \bar{\mathbf{x}}) = \int p(\bar{\boldsymbol{\theta}}|J) \prod_k p(\mathbf{z}_{kf}|\boldsymbol{\theta}_{kf}) p(\mathbf{z}_{kb}|\boldsymbol{\theta}_{kb}) d\bar{\boldsymbol{\theta}} \ . \quad (1)$$

Due to the number of pixels, the likelihood will be sharp so for simplicity we approximate the integral over $\boldsymbol{\theta}$ with the maximum[1]:

$$p(\bar{\mathbf{z}}|J = 0, \bar{\mathbf{x}}) \approx \max_{\bar{\boldsymbol{\theta}}} p(\bar{\boldsymbol{\theta}}) \prod_k p(\mathbf{z}_{kf}|\boldsymbol{\theta}_{kf}) p(\mathbf{z}_{kb}|\boldsymbol{\theta}_{kb}) \quad (2a)$$

$$p(\bar{\mathbf{z}}|J = 1, \bar{\mathbf{x}}) \approx \max_{\boldsymbol{\theta}_{1f} = \boldsymbol{\theta}_{2f}} p(\bar{\boldsymbol{\theta}}) \prod_k p(\mathbf{z}_{kf}|\boldsymbol{\theta}_{kf}) p(\mathbf{z}_{kb}|\boldsymbol{\theta}_{kb}) \ . \quad (2b)$$

Under this approximation, $p(\bar{\mathbf{z}}|J = 0, \bar{\mathbf{x}}) \geq p(\bar{\mathbf{z}}|J = 1, \bar{\mathbf{x}})$ always.

We want to choose the segmentations $\mathbf{x}_k$ so that the hypothesis $J = 1$ has high posterior probability. In other words, we want to find

$$\bar{\mathbf{x}}^\star = \operatorname*{argmax}_{\bar{\mathbf{x}}} p(J = 1|\bar{\mathbf{z}}, \bar{\mathbf{x}}) p(\bar{\mathbf{x}}) \quad (3)$$

[1]This approximation does not affect the final answer up to constant [18].

where $p(J=1|\bar{\mathbf{z}},\bar{\mathbf{x}}) =$

$$\frac{p(\bar{\mathbf{z}}|J=1,\bar{\mathbf{x}})p(J=1)}{p(\bar{\mathbf{z}}|J=0,\bar{\mathbf{x}})p(J=0)+p(\bar{\mathbf{z}}|J=1,\bar{\mathbf{x}})p(J=1)} .$$

We will set $p(J=0)=p(J=1)$, so these terms disappear. To simplify the formula, define

$$D(\bar{\mathbf{z}}|\bar{\mathbf{x}}) = \log \frac{p(\bar{\mathbf{z}}|J=0,\bar{\mathbf{x}})}{p(\bar{\mathbf{z}}|J=1,\bar{\mathbf{x}})} . \quad (4)$$

In this ratio, the background terms cancel, and we will obtain a measure of divergence between the foreground areas of $\mathbf{z}_1$ and $\mathbf{z}_2$. Under the approximation (2), $D \geq 0$.

Taking the negative logarithm of (3) gives the following energy minimization problem:

$$\bar{\mathbf{x}}^\star = \operatorname*{argmin}_{\bar{\mathbf{x}}} \log(1+\exp(D(\bar{\mathbf{z}}|\bar{\mathbf{x}}))) - \log p(\bar{\mathbf{x}}) \quad (5a)$$

$$\approx \operatorname*{argmin}_{\bar{\mathbf{x}}} \frac{1}{2} D(\bar{\mathbf{z}}|\bar{\mathbf{x}}) - \log p(\bar{\mathbf{x}}) . \quad (5b)$$

This approximation is justified when D is small at the optimum.

Prior $p(\mathbf{x})$ We use an MRF model for each image. Furthermore, we assume that larger foreground regions are more likely a priori. Thus, we have

$$-\log p(\bar{\mathbf{x}}) = \lambda_{bg} \sum_{k,i} (1-x_{ki}) + \sum_{k,(i,j)} \lambda_{ki,kj} |x_{ki}-x_{kj}| + const \quad (6)$$

where the second sum is over pairs of neighboring pixels. We use the following expression for coefficients $\lambda_{ki,kj}$:

$$\lambda_{ki,kj} = \lambda_1 + \lambda_2 \exp(-\beta ||I_{ki}-I_{kj}||^2)$$

where I_{ki} is the colour of pixel i in image k and $\beta = (2\left\langle ||I_{ki}-I_{kj}||^2 \right\rangle)^{-1}$. This is similar to the contrast-sensitive term in [17], with the addition of Ising prior λ_1.

Image generation model The remaining task is to specify the image generation model for the foreground region: $p(\mathbf{z}_{kf}|\boldsymbol{\theta}_{kf})$. By choosing this model carefully, we obtain a commonly used divergence measure D. Our choice is a Gaussian model on histograms, which leads to the classical variational distance. Let $\hat{h}_{kf}$ be the empirical un-normalized histogram of foreground pixels: $\hat{h}_{kf}(z) = \sum_i x_{ki}\delta(z_{ki}-z)$. Given a histogram, the foreground is generated by laying out exactly the expected number of pixels from each colour bin, then randomly permuting them. Therefore $p(\mathbf{z}_{kf}|\boldsymbol{\theta}_{kf})$ is proportional to $p(\mathbf{h}_{kf}=\hat{h}_{kf}|\boldsymbol{\theta}_{kf})$ and we only need to specify the distribution $p(\mathbf{h}_{kf}|\boldsymbol{\theta}_{kf})$. In the following, everything concerns foreground so we will drop the f subscript. The target histogram $\mathbf{h}_k$ is generated by a Gaussian distribution with parameters $\boldsymbol{\theta}_k = (\mathbf{m}_k, \mathbf{v}_k)$, with hyperparameter c_k controlling the expected size of the foreground region:

$$p(\mathbf{h}_k|\mathbf{m}_k,\mathbf{v}_k) = \prod_z \mathcal{N}(h_k(z); c_k m_k(z), c_k^2 v_k(z)) . \quad (7)$$

Note that $(\mathbf{m}_k, \mathbf{v}_k)$ are shared under $J=1$ but c_k is not. Therefore c_k can compensate for foreground size differences among the images. We will use a uniform prior on $\mathbf{m}_k$ and Gamma prior on $\mathbf{v}_k$:

$$p(v_k(z)) \propto v_k(z) \exp(-\frac{v_k(z)}{b^2}) . \quad (8)$$

When $J=0$, $\hat{\mathbf{m}}_k = \hat{h}_k/c_k$ and $\hat{\mathbf{v}}_k = \mathbf{0}$ so:

$$p(\bar{\mathbf{z}}|J=0,\bar{\mathbf{x}}) \approx 1 . \quad (9)$$

When $J=1$:

$$\hat{\mathbf{m}}_\star = \frac{1}{2}\sum_k \frac{\hat{h}_k}{c_k} \quad , \quad \hat{v}_\star(z) = \frac{b}{2}\left|\frac{\hat{h}_1(z)}{c_1} - \frac{\hat{h}_2(z)}{c_2}\right| \quad (10)$$

$$p(\bar{\mathbf{z}}|J=1,\bar{\mathbf{x}}) \approx \prod_z p(\hat{v}_\star(z)) \prod_k \mathcal{N}(\hat{h}_k(z); \hat{\mathbf{m}}_\star(z), \hat{v}_\star(z))$$

$$D(\bar{\mathbf{z}}|\bar{\mathbf{x}}) = \frac{1}{b}\sum_z \left|\frac{\hat{h}_1(z)}{c_1} - \frac{\hat{h}_2(z)}{c_2}\right| . \quad (11)$$

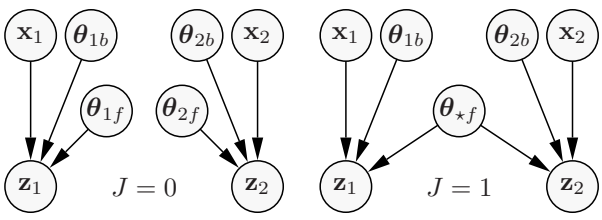

Figure 2. **The two hypotheses for image generation.**

3. Optimization

In the previous section we described a generative model that yields the following energy function:

$$E(\bar{\mathbf{x}}; c_1, c_2) = -\log p(\bar{\mathbf{x}}) + E^{global}(\hat{h}_1, \hat{h}_2; c_1, c_2) \quad (12)$$

The first term is given by (6); it encodes the usual MRF prior on labeling $\mathbf{x}$. The second term is quite different from the first one: it depends on *global* properties of segmentation $\mathbf{x}$, namely histograms of foreground regions $\hat{h}_1, \hat{h}_2$:

$$E^{global}(\hat{h}_1, \hat{h}_2; c_1, c_2) = \frac{1}{2b}\sum_z \left|\frac{\hat{h}_1(z)}{c_1} - \frac{\hat{h}_2(z)}{c_2}\right| . \quad (13)$$

The presence of this global term makes the minimization problem quite challenging. One could think of using some general inference algorithm, such as Swendsen-Wang Cuts for sampling arbitrary posterior probabilities [1]. Another

possibility is to use active contours [12, 9]. We argue, however, that since the MRF term is an essential part of the energy, it is desirable to use the well-established technique for binary MRFs - min cut/max flow algorithm [4]. Fortunately, the form of our global term will allow to use max flow algorithm inside the method called *submodular-supermodular procedure* [15].

For simplicity, in this paper we set $c_1 = c_2 = 1$, which means that we prefer foreground regions of the same size. It is easy, however, to extend the model to account for different sizes: we can put some prior on c_1, c_2 and minimize energy (12) iteratively, i.e. fix c_1, c_2 and optimize over $\bar{\mathbf{x}}$ and then the other way around.

We now describe how we minimize energy (12). We iterate between the following two steps: (1) Fix $\mathbf{x}_2$, optimize over $\mathbf{x}_1$, and (2) fix $\mathbf{x}_1$, optimize over $\mathbf{x}_2$. Each subproblem requires minimizing the following function:

$$-\log p(\mathbf{x}_k) + \frac{1}{2b}\sum_z |\hat{h}_k(z) - h^{target}(z)| \qquad (14)$$

where the target histogram h^{target} is the empirical histogram of the foreground in other image. For the remainder of this section we focus on how to solve this subproblem for a given image k. Since k is fixed, we will omit it for brevity.

The energy function can be written as

$$E(\mathbf{x}) = E_1(\mathbf{x}) + E_2(\mathbf{x}) \qquad (15)$$

where the first term corresponds to the prior on $\mathbf{x}$ and the second to the global histogram term. An important observation is that E_1 is *submodular* and E_2 is *supermodular*, i.e. they satisfy

$$E_1(\mathbf{x} \wedge \mathbf{x}') + E_1(\mathbf{x} \vee \mathbf{x}') \leq E_1(\mathbf{x}) + E_1(\mathbf{x}')$$
$$E_2(\mathbf{x} \wedge \mathbf{x}') + E_2(\mathbf{x} \vee \mathbf{x}') \geq E_2(\mathbf{x}) + E_2(\mathbf{x}')$$

for all configurations $\mathbf{x}, \mathbf{x}'$.

It is well-known that any submodular function can be minimized in polynomial time [19]. In our case $E_1(\mathbf{x})$ is a sum of unary and pairwise terms, so a global minimum of E_1 can be computed very efficiently via min cut/max flow algorithm. The presence of supermodular part, however, makes the problem NP-hard.

The submodular-supermodular procedure [15] is a promising approximate minimization technique for functions of the form (15). Sec. 3.1 gives an overview of this approach. Sec. 3.2 we discusses its potential difficulties and proposes an alternative method - *trust region graph cuts*.

3.1. Submodular-supermodular procedure (SSP)

This method was inspired by concave-convex procedure for minimizing functions of continuous variables [22]. SSP is an iterative technique which produces a sequence of configurations $\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^t, \ldots$. The main property of SSP is that the energy never goes up, i.e. $E(\mathbf{x}^0) \geq E(\mathbf{x}^1) \geq \ldots$. Let $\mathbf{x}^t$ be the current configuration. The method performs the following steps:

(a) Replace supermodular part $E_2(\mathbf{x})$ with a linear approximation $\widehat{E}_2(\mathbf{x}) = C + \langle \mathbf{x}, \mathbf{y} \rangle = C + \sum_i x_i y_i$ where C is a constant and $\mathbf{y}$ is a real-valued vector. (Such a function is also called *modular*).

(b) Compute a global minimum of function $E_1(\mathbf{x}) + \widehat{E}_2(\mathbf{x})$ to get new configuration $\mathbf{x}^{t+1}$.

Note that minimization in the second step can be performed in polynomial time since the function is submodular. (Linear function $\langle \mathbf{y}, \mathbf{x} \rangle$ simply adds unary terms to $E_1(\mathbf{x})$).

Linear approximation chosen in step (a) must satisfy two properties: (i) It must be an upper bound on the supermodular part, i.e. $\widehat{E}_2(\mathbf{x}) \geq E_2(\mathbf{x})$ for all configurations $\mathbf{x}$. (ii) The functions should touch at $\mathbf{x}^t$: $\widehat{E}_2(\mathbf{x}^t) = E_2(\mathbf{x}^t)$. These properties ensure that the original energy does not go up since $E(\mathbf{x}^{t+1}) \leq E_1(\mathbf{x}^{t+1}) + \widehat{E}_2(\mathbf{x}^{t+1}) \leq E_1(\mathbf{x}^t) + \widehat{E}_2(\mathbf{x}^t) = E(\mathbf{x}^t)$.

It remains to specify how to choose an upper bound $\widehat{E}_2(\mathbf{x})$ (i.e. corresponding vector $\mathbf{y}$) with the properties above. (Existance of such a bound follows from supermodularity of E_2). [15] uses the following procedure. First, an ordering of nodes $\pi(\cdot)$ is selected which "respects" current labeling $\mathbf{x}^t$, i.e. all ones precede all zeros: $x^t_{\pi(1)} \geq x^t_{\pi(2)} \geq \ldots \geq x^t_{\pi(n)}$. This ordering defines the following $n+1$ configurations: $\mathbf{x}^{(0)} = (0, 0, \ldots, 0), \mathbf{x}^{(1)} = (1, 0, \ldots, 0), \ldots, \mathbf{x}^{(n)} = (1, 1, \ldots, 1)$, where we assumed that the nodes are ordered according to π. (Formally, $x_i^{(j)}$ is zero if $\pi(i) \leq j$, and one otherwise). The fact that ordering π "respects" current labeling $\mathbf{x}^t$ simply means that $\mathbf{x}^t$ is one of these $n+1$ configurations. Finally, approximation $\widehat{E}_2(\mathbf{x})$ is chosen so that it is exact for these $n+1$ configurations: $\widehat{E}_2(\mathbf{x}^{(j)}) = E_2(\mathbf{x}^{(j)})$, $j = 0, 1, \ldots, n$. Solving $n+1$ equations with $n+1$ unknowns yields

$$C = E_2(\mathbf{x}^{(0)}) \ , \ y_{\pi(i)} = E_2(\mathbf{x}^{(i)}) - E_2(\mathbf{x}^{(i-1)})$$

3.2. Trust region graph cuts (TRGC)

For SSP it is important to choose "good" representative configurations $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(n)}$. If, for example, a global minimum of $E(\cdot)$ happens to be among these configurations, then the procedure will find this minimum. Choosing good configurations, however, is a difficult problem. First, there is a restriction on representative configurations[2]: there must

[2]Note that this restriction on representative configurations does not necessarily mean that SSP cannot "exchange" pixels. If some configuration is not among $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(n)}$, it may still happen that approximation $\widehat{E}_2(\mathbf{x})$ is tight for this configuration. Furthermore, even if the approximation is not very tight, theoretically it is still possible that SSP will go there.

hold either $\mathbf{x}^{(j)} \leq \mathbf{x}^t$ or $\mathbf{x}^{(j)} \geq \mathbf{x}^t$. Second, even if there is an ordering that would decrease the energy, computing such an ordering is an NP-complete problem (see discussion in [18]).

It could be desirable to choose linear approximation $\widehat{E}_2(\mathbf{x}) = C + \langle \mathbf{x}, \tilde{\mathbf{y}} \rangle$ which is not based on any ordering. For example, we could set $\tilde{y}_i = E_2(\mathbf{x}^{i;1}) - E_2(\mathbf{x}^{i;0})$ where $\mathbf{x}^{i;s}$ is the labeling obtained from $\mathbf{x}^t$ by setting x_i to s. This approximation is exact for all configurations that differ from $\mathbf{x}^t$ by at most one pixel. It can also be obtained by keeping linear terms in the Taylor expansion of energy E_2 expressed as a function of the global histogram of $\mathbf{x}$ (assuming that E_2 is differentiable).

Unfortunately, this approximation is not an upper bound on $E_2(\mathbf{x})$. This means that minimizing $E_1(\mathbf{x}) + \widehat{E}_2(\mathbf{x})$ is not guaranteed to decrease the original energy. To remedy this problem, we propose an alternative method which we call *trust region graph cuts (TRGC)*. It allows arbitrary linear approximations $\widehat{E}_2(\mathbf{x})$ which are not upper bounds. Furthermore, in this method function $E_2(\mathbf{x})$ can also be arbitrary - it is no longer required to be supermodular.

Trust region methods are well-known in continuous optimization [2]; TRGC can be viewed as their discrete analogue. A related continuous optimization method is the linearization method of Pschenichnyj [16].

Description of TRGC Instead of selecting unary potentials $\mathbf{y}$ based on some ordering, we will *optimize* over $\mathbf{y}$. Our technique produces a sequence of vectors $(\mathbf{x}^0, \mathbf{y}^0), \ldots, (\mathbf{x}^t, \mathbf{y}^t), \ldots$ with the following properties: (i) $\mathbf{x}^t = \arg\min_{\mathbf{x}} E_1(\mathbf{x}) + \langle \mathbf{x}, \mathbf{y}^t \rangle$, and (ii) the energy does not go up: $E(\mathbf{x}^0) \geq E(\mathbf{x}^1) \geq \ldots$.

The method works as follows. Let $(\mathbf{x}^t, \mathbf{y}^t)$ be the current state, and $\widehat{E}_2(\mathbf{x}) = C + \langle \mathbf{x}, \tilde{\mathbf{y}} \rangle$ be a desired approximation of $E_2(\mathbf{x})$. Let us define $\mathbf{y}(\alpha) = (1-\alpha)\mathbf{y}^t + \alpha\tilde{\mathbf{y}}$, and let $\mathbf{x}(\alpha)$ be a global minimum of function $E_1(\mathbf{x}) + \langle \mathbf{x}, \mathbf{y}(\alpha) \rangle$[3]. Note that $\alpha = 0$ corresponds to the current solution $\mathbf{x}^t$, and $\alpha = 1$ corresponds to taking approximation $\widehat{E}_2(\mathbf{x})$. We now search for $\alpha \in [0, 1]$ that minimizes $E(\mathbf{x}(\alpha))$. This defines new vectors $\mathbf{x}^{t+1}$ and $\mathbf{y}^{t+1}$. If $\alpha = 0$ is within the range of values that we test, the energy is guaranteed not to go up.

We implemented the following one-dimensional search routine: we start with $\alpha = 1$ and we keep halving it until one of the following happens: (a) $\mathbf{x}(\alpha) = \mathbf{x}^t$; (b) $\alpha < 10^{-3}$; or (c) energy $E(\mathbf{x}(\alpha))$ is larger compared to the previous α, and the energy for the previous α was smaller than $E(\mathbf{x}^t)$.

It is important to note that TRGC is a trust region method working in the *dual* space: we optimize over dual variables $\mathbf{y}$ rather than primal variables $\mathbf{x}$.

[3] If there are multiple global minima, then $\mathbf{x}(\alpha)$ will denote one of the them. There is one exception, however: if $\mathbf{x}^t$ is also a global minimum, then we set $\mathbf{x}(\alpha) = \mathbf{x}^t$.

3.3. Implementational details

The general structure of the algorithm for cosegmenting an image pair is described in the beginning of sec. 3. The remaining question is the initialization of the target distributions and the segmentation for the first iteration. For this we employ a procedure which finds the largest regions in two images of the same size whose histograms match perfectly. This is done via a greedy algorithm that adds one pixel at a time to the first and second foreground regions. Note, this gives the minimum energy if the spatial prior is ignored.

SSP. The most important question for SSP is how to choose an ordering of nodes π. We tested two schemes. In the first one we selected a random permutation of elements that respects current configuration $\mathbf{x}$. This is similar to the technique used in [15], with one modification: we take random permutation of 10×10 blocks rather than individual pixels. Inside each block pixels with the same segmentation are ordered sequentially. Thus, we try to take into account the fact that due to spatial coherence all pixels inside a block are likely to have the same segmentation. Our second scheme is deterministic: given initial configuration, we compute a signed distance map from segmentation boundary and order pixels according to this distance. In this scheme representative configurations $\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(n)}$ correspond to diluting or eroding the current foreground region. For a fixed target histogram we ran a maximum of 50 iterations of SSP procedure. We observed, however, that in the majority of cases only the first few iterations decrease the energy, and then the energy stays constant.

TRGC. We used the SSP procedure for initialization (i.e. for computing $(\mathbf{x}^0, \mathbf{y}^0)$). We ran the algorithm until convergence, i.e. until searching over α did not yield any improvement in the energy.

For both approaches we used maxflow algorithm in [4]. Furthermore for all experiments we set $\lambda_{bg} = 0.3$, $\lambda_1 = 1, \lambda_2 = 50$ and $b = 0.5$. Finally let us introduce our appearance model. We have experimented with a simple 2D intensity normalized RGB colour space and a richer texture (texton) based model [14], which has been proven to be very powerful for image retrieval [7]. Apart from scenarios of retrieving images of the same class we have used the simple model since the emphases on colour improved the performance, if the common part is the identical object. A thorough testing of different appearance models is a part of future work.

4. Experiments

Comparison of SSP and TRGC. We built a data set of 50 images which depict a foreground object in front of a background. The ground truth segmentation of the foreground object has been achieved manually[4]. In some im-

[4] The data set is publicly available at http://research.microsoft.com/vision/cambridge/i3l/segmentation/GrabCut.htm

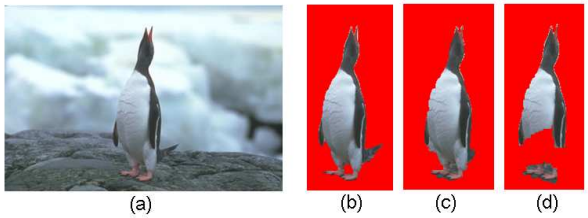

Figure 3. **Comparison of TRGC and SSP.** The goal is to segment the object (penguin) in the input image (a) given the target histogram of the ground truth segmentation (b). The result of TRGC (c) clearly outperforms SSP (d).

Method	av. Energy	av. Error (%)	av. # Iter.
TRGC(dist.)	408	2.33	7.8
TRGC (rand.)	417	2.33	7.8
SSP (dist.)	426	2.77	4.6
SSP (rand.)	461	2.81	4.6
Ground Truth	429	0.0	-

Table 1. **Comparison of SSP and TRGC** with radome ordering (rand.) and distance map ordering (dist.) of the nodes. Note that the energies are scaled by 10^{-2}.

ages the object is "camouflaged" (e.g. fig. 3(left)), where fore- and background have similar appearance, in other images (e.g. 4(left)) they have different appearances. Given the target histogram of the ground truth segmentation we compare the performance of the submodular-supermoduler procedure (SSP) with our version (TRGC). We also compare the performance of ordering of the nodes (see sec. 3.3): Random ordering (rand.), as suggested in [15], versus distance map ordering (dist.). As performance measure we utilize the average energy (av. Energy) and the percentage of misclassified pixels (av. Error) with respect to ground truth. The results are summarized in table 4. It is clear that TRGC outperforms SSP considerably both in terms of lower energy and quality of result. Note that the energy of TRGC was *always* lower than SSP. With respect to the pixel ordering: random versus distance transform, the later performs slightly better, and is also deterministic. Consequently we used the TRGC method with distance transform ordering for initialization as our method for the remaining experiments. Fig. 3 shows an example where TRGC outperforms SSP. Note, the fact that the ground truth has a relatively low energy shows that our problem setting is reasonable.

Examples of cosegmentation using TRGC are shown in fig. 1,4-7. Fig. 4 demonstrates that the segmentation quality depends on the background penalty λ_{bg}. Our generative framework gives us the option of learning this parameter given a training and validation data set. To obtain such a database is part of future work.

Robust Image distance for Image retrieval. In the following we consider two examples where we demonstrate that cosegmentation improves an image retrieval system based on global histogram comparison. The key idea is to use the energy as a distance measure between an image pair. This is a valid measurement since identical images have energy (distance) 0. Another nice feature of our energy is that by adjusting $\lambda_{bg} = \infty$ it gives the standard global histogram difference of the whole image, as used in e.g. [7]. As in all previous examples we set $\lambda_{bg} = 0.3$.

In fig. 5 we compare the distance between three images where two of them depict the same scene and the third an unrelated scene. We demonstrate that using cosegmentation two images of the same scene have a smaller distance than two unrelated images. This is in contrast to using an appearance statistics of the whole image where two unrelated images have a smaller distance (details in figure caption).

In the second example, fig. 6, we compare the distance of a triplet of images where two images depict an object of the same class (bus) and a third unrelated image. The findings are as in the previous case, cosegmentation gives the correct relationship for the triplet (see figure caption for details). Given the middle image in fig. 6 as query, the right image is in fact the most similar image from the Corel database of 1000 images used in [11] and based on global texture (texton [14]) statistics. The fact that our cosegmentation system returns an image containing an object of the same class (fig. 6 left) is a proof of concept that the retrieval performance for this particular query image improves. Further quantitative tests on the whole database have to be carried out. In particular, it has to be tested that ignoring the similarity of the background does not decrease performance for a query image which does not contain a well defined object.

Further Applications. Let us demonstrate other applications where our generative framework can be applied successfully. Fig. 7 shows an example for video summarization and interactive cosegmentation (see figure caption for details). Fig. 8 depicts an application where our generative framework is used for automatically tracking and segmenting a foreground object in a video sequence given a target distribution in the first key frame (details in figure caption).

5. Conclusion and Future Work

We have presented a novel generative model for cosegmenting the common parts of an image pair. The strength of the model is its generality: The common part can be a rigid/non-rigid object (or scene), observed from different viewpoints or even similar objects of the same class. Inference in the model leads to minimization an energy with an MRF term encoding spatial coherency and a global constraint which tries to match the appearance histograms of the common parts. This exact energy has not been proposed earlier and its optimization is challenging and NP-hard. We have presented a novel optimization scheme which we call trust region graph cuts, and have demonstrated its superiority to a competitive method on a large data set. Our new framework has clear applications for interactive graphics,

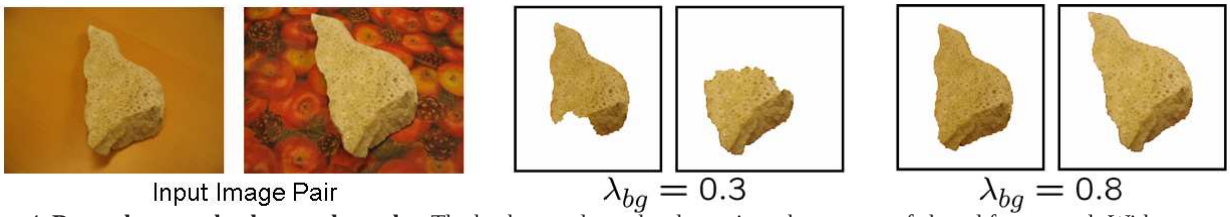

Figure 4. **Dependency on background penalty.** The background penalty determines the amount of shared foreground. With our standard setting of $\lambda = 0.3$ only part of the object was detected. By increasing $\lambda_{bg} = 0.8$ we force more foreground material to appear. Given our generative model we plan to learn λ_{bg} from a larger training/validation data set.

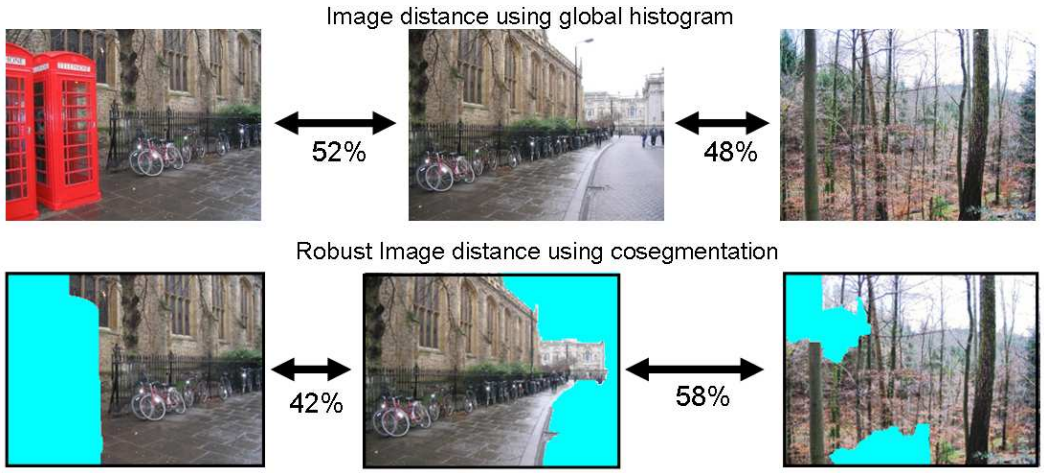

Figure 5. **Robust Image distance - same scene.** Consider the triplet of images in the top row. The left and middle image depict part of the same scene, where the right image shows an unrelated forest scene. The distance (SAD) of the global colour histograms of the whole images says that the middle image is more similar to the right (48%) than to the left image (52%). Running cosegmentation gives the expected answer (bottom row). The cosegmentation of the left and middle image nicely moves the regions which do not appear in both images (telephone box, sky and road) to the background (label light blue). Note that the depicted cosegmentation of the middle image is with respect to the left image. When using the energy of the cosegmentation as distance measure, the middle image is now more similar (42%) to the left than the right image (58%). Note that the percentages are derived by comparing the absolute energies. Also, note that the cosegmentation measure without the spatial coherence term (MRF) gives, as the global histogram of whole images, the incorrect answer.

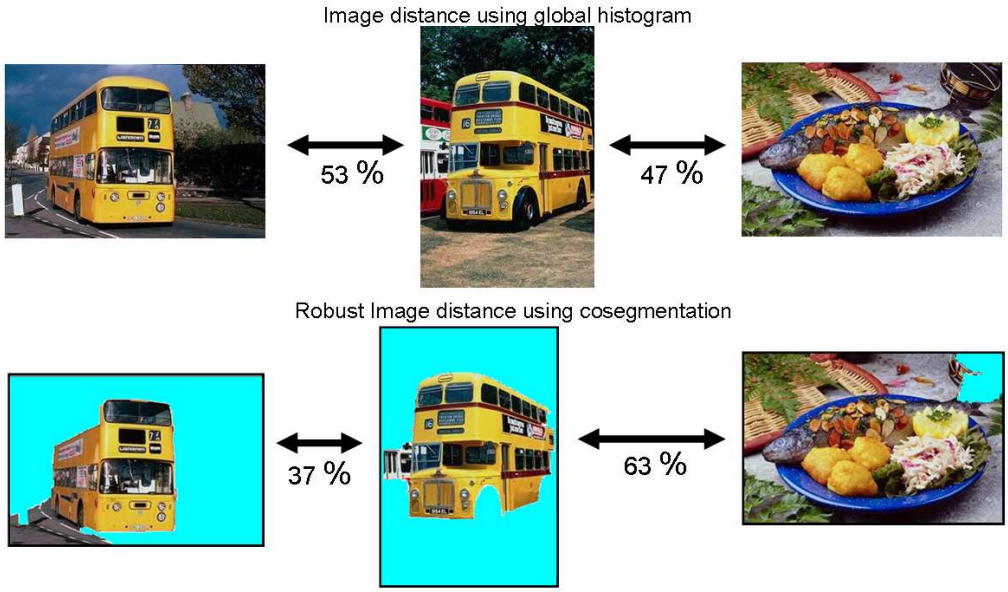

Figure 6. **Robust Image distance - similar objects.** Same explanation as in fig. 5, apart from the fact that the appearance model is based on texture (textons [14]). Note that the trees in the background where assigned a different texton label.

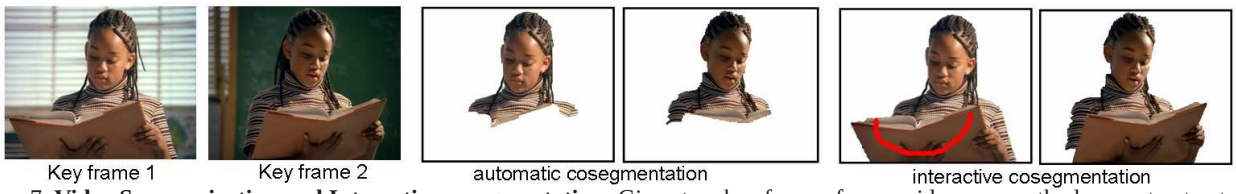

Figure 7. **Video Summarization and Interactive cosegmentation.** Given two key frames from a video, our method can extract automatically the common part. This can be used to summarize the video. In this case the segmentation is not perfect, due to colour variations on the book cover. In an interactive cosegmentation system the foreground object can be extracted from *both* images, by editing only *one* image. We utilize the interactive brushing style of [3]. In the image (second from right) a red brush stroke indicates an explicit marking of the foreground. Obviously, the updated histogram of the left image forced a better solution for the right image.

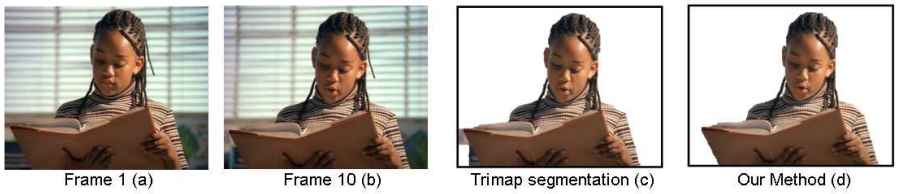

Figure 8. **Video Tracking and Segmentation.** Given a perfect segmentation in a key frame (a) we would like to segment the the foreground object in all subsequent frames, e.g. fame 10 (b). An obvious solution is to apply standard image segmentation [3] using a trimap, which is derived by dilating the segmentation of the previous frame by a fixed number of pixels. The result (c) is good, however the segmentation of the book is sub-optimal. Our result (d) is better, by forcing the foreground object to have the same target histogram as in the previous frame.

video tracking and segmentation. Probably the most important application is object-driven image retrieval, for which we propose a new and robust similarity measurement for image pairs. In the future we hope to quantify our initial findings in this area. A further future direction is the incorporation of feature matches (optical flow) which is an essential component of any standard wide-baseline matching, or tracking system. Also a comparison with an alternative generative model, introduced in [18], is important.

References

[1] A. Barbu and S. C. Zhu. Generalizing Swendsen-Wang to sampling arbitrary posterior probabilities. *PAMI*, 27, 05.

[2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

[3] Y. Boykov and M.-P. Jollie. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, 2001.

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9), September 2004.

[5] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *ICCV*, pages 1197–1203, 1999.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *PAMI*, 25(5):564–575, 2003.

[7] T. Deselaers, D. Keysers, and H. Ney. Classification error rate for quantitative evaluation of content-based image retrieval systems. In *ICPR*, 2004.

[8] D. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Optical Soc. of America A.*, 4:2379–2394, 1987.

[9] D. Freedman and T. Zhang. Active contours for tracking distributions. *IEEE T. Image Processing*, 13(4), 2004.

[10] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact MAP estimation for binary images. *J. Royal Statistical Society*, 51:271–279, 1989.

[11] Z. James, J. l. Wang, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *PAMI*, 23(9):947–963, 2001.

[12] S. Jehan-Besson, M. Barlaud, G. Aubert, and O. Faugeras. Shape gradients for histogram segmentation using active contours. In *ICCV*, 2003.

[13] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Bi-layer segmentation of binocular stereo video. In *CVPR*, 2005.

[14] T. Leung and J. Malik. Recognizing surfaces using three-dimensional textons. In *ICCV*, 1999.

[15] M. Narasimhan and J. Bilmes. A supermodular-submodular procedure with applications to discriminative structure learning. In *UAI*, July 2005.

[16] B. N. Pshenichnyj. *The Linearization Method for Constrained Optimization*. Computat. Mathematics 22. 1994.

[17] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *SIGGRAPH*, 23(3):309–314, 2004.

[18] C. Rother, V. Kolmogorov, T. Minka, and A. Blake. Cosegmentation of image pairs by histogram matching — incorporating a global constraint into MRFs. Technical Report MSR-TR-2006-36, 2006.

[19] A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *J. of Combinatorial Theory, Ser. B*, 80:346–355, 2000.

[20] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 2000.

[21] J. Winn and N. Jojic. LOCUS: Learning object classes with unsupervised segmentation. In *ICCV*, 2005.

[22] A. L. Yuille and A. Rangarajan. The concave-convex procedure (CCCP). In *NIPS*, 2001.