IEEE Conference on Computer Vision and Pattern Recognition 2012 (CVPR'12)

Linear Solution to Scale Invariant Global Figure Ground Separation

Hao Jiang
Computer Science Department, Boston College, USA

Abstract

We propose a novel linear method for scale invariant figure ground separation in images and videos. Figure ground separation is treated as a superpixel labeling problem. We optimize superpixel foreground and background labeling so that the object foreground estimation matches model color histogram, its area and perimeter are consistent with object shape prior, and the foreground superpixels form a connected region. This optimization problem is challenging due to high-order soft and hard global constraints among large number of superpixels. We devise a scale invariant linear method that gives an integer solution with a guaranteed error bound via a branch and cut procedure. The proposed method does not rely on motion continuity and works on static images and videos with abrupt motion. Our experimental results on both synthetic ground truth data and real images show that the proposed method is efficient and robust over object appearance changes, large deformation and strong background clutter.

1. Introduction

We propose a novel method for scale invariant global figure ground separation. We assume that the foreground properties are known, e.g., the object color histogram and shape description are available, but the background properties are are unknown. This is a typical scenario when we localize objects in images and videos. We also do not assume motion continuity so that the proposed method can be used for foreground estimation in static images and challenging videos. Fig. 1 illustrates the problem we tackle.

We treat figure ground separation as a labeling problem. We label over-segmented superpixels in images as foreground and background so that the foreground region is consistent with the object model in appearance and shape, and at the same time the foreground superpixels are connected. This is a challenging problem due to the high order coupling among large number of superpixels. To our knowledge this global segmentation problem has not been well explored. It is an open question whether an efficient solution exists to satisfy all the global constraints and is able to achieve global optimal or near global optimal result.

In this paper, we propose an efficient linear method to tackle the global figure ground separation problem. We construct a scale invariant linear formulation and a branch and cut procedure to directly obtain the integer solution. We apply the proposed method in localizing both rigid and highly deformable objects in images and videos. The proposed method has the following properties: (1) It is scale invariant, (2) satisfies global appearance, shape and connectivity constraints, and (3) it efficiently solves the global segmentation problem with an error bound guarantee.

Figure 1. Using foreground model extracted from one image (upper left) we find the object foregrounds in images using a scale invariant linear approach. The foreground estimation follows model histogram, shape and maintains region connectivity.

Object figure ground separation has been intensively studied. One popular scheme is to extract a moving object by tracking. By combining tracking and segmentation, human subjects can be reliably localized in videos [1] using Conditional Random Field and Belief Propagation. In [2], pixel level tracking and foreground labeling are optimized simultaneously in a dual decomposition framework. Video segmentation [7] has also been studied to extract 3D superpixels. Automatic human segmentation [3] has been achieved by combining people detection, level set method and Belief Propagation to link foreground estimations into a continuous volume. Unsupervised method [6] has recently been proposed to group object category independent proposals [5] into object trajectories. These previous methods achieve accurate results on videos with continuous motion but cannot be easily extended to process videos with motion discontinuity. We propose a new method to tackle the problem. The proposed method also does not reply on specific object detectors and can thus be directly applied to many different objects.

Foreground connectivity is a strong prior for figure ground separation. Connected Markov Random Field [11] and Steiner tree method [4] [15] have been successfully

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

applied to connectivity constrained figure ground separation in object class segmentation. These previous methods use unary and pairwise terms in the objective function and require both foreground and background models. Ideally foreground superpixels have mostly negative costs and background superpixels have mostly positive costs. Foreground object can thus be extracted by finding connected superpixels that have minimum total cost. This scheme cannot be easily extended to solve our problem because we only have the foreground model and our local costs all have the same sign. Minimizing the overall cost would give a trivial result. A foreground size constraint can be included to relieve the problem. However, this still does not solve the problem because the composition of features is not considered in these previous methods. Using only unary and pairwise objective would cause segmentation error if the background clutter has similar appearance to the foreground object. We need a new formulation that includes high order terms to constrain the overall object appearance, shape, connectivity, and the formulation needs to be scale invariant.

Including global constraints such as color histogram consistency in the optimization improves the segmentation result but at the same time makes the problem hard to solve. Different approximation methods have been proposed. In [9], image cosegmentation is first proposed. This formulation uses an L_1 norm color histogram consistency term, and an iterative graph-cut approach is proposed to obtain an approximation solution to the linear combination of the sub- and super-modular problem. Another iterative graph-cut method using Bhattacharyya distance to measure color histogram difference is proposed in [17] for scale invariant figure ground separation. Foreground connectivity is not explicitly constrained in these methods. Convex relaxation methods have also been intensively studied. Linear relaxation for a quadratic formulation [10] is proposed to solve the cosegmentation problem. Discriminative clustering and quadratic optimization have also been successfully applied to cosegmentation in [8]. In [13], linear relaxation method with spatial constraints has been proposed to group superpixels into object foreground. With a training set, object part interaction in the segmentation hierarchy can also be extracted for robust object class segmentation [12] using non-convex quadratic programming. Relaxation methods yield floating point results and still need to be converted to integer solutions and these previous methods do not explicitly constrain the foreground continuity.

The contribution of this paper is a scale invariant linear method for global figure ground separation, which efficiently yields integer solutions with a guaranteed error bound. The proposed method only uses the object foreground model. It models both global soft and hard constraints so that the estimation complies with the model appearance and shape, and it guarantees the connectivity of the foreground estimation. These properties are critical for reliable segmentation results.

2. Method

Global figure ground separation can be written as the following optimization problem.

$$\min_{x,s}\{H(x,s)+\lambda A(x,s)+\mu P(x,s)+\gamma T(x)\}$$
$$\text{s.t. Foreground is connected under labeling } x. \qquad (1)$$

Here, x is the foreground/background superpixel labeling and s is an unknown scale. $H(x,s)$ measures the similarity between the foreground histogram and the model; H is small if the labeled foreground region fits the model histogram at a specific scale s. $A(x,s)$ and $P(x,s)$ quantify the area and perimeter difference between the model and the labeled foreground region at scale s. A and P terms thus constrain the shape of the foreground estimation. $T(x)$ includes the optional unary and pairwise terms. λ, μ, γ are constant parameters controlling the weight among different terms. In this paper, $\lambda = \mu = 1$ and $\gamma = 0.01$. Apart from the soft constraints on x, the labeling has to yield a connected foreground estimation. Eq. (1) defines a challenging combinatorial optimization problem. Naive exhaustive search is not an option. In the following, we linearize the problem so that we can efficiently find its lower bound and use it as a basis for an efficient branch and cut procedure.

2.1. Linearization

We label each superpixel as foreground or background in the target image. Let x_i be the indicator variable for superpixel $i = 1..N$. If superpixel i is on the foreground $x_i = 1$, and if it is on the background $x_i = 0$. The unary cost of the superpixel labeling is thus $\sum_{i=1}^{N} c_i x_i$, where c_i measures how well superpixel i matches the foreground model. Note that we only have the object foreground model and $c_i \geq 0$ for each i. Directly minimizing the unary cost would thus give a trivial all 0 solution. We use a high order formulation to solve the problem.

2.1.1 Histogram Matching

We enforce that the overall color histogram of the foreground estimation should be close to the model color histogram. Let $\tilde{h}(k)$ be the model color histogram at bin k and $h_i(k)$ be superpixel i's color histogram at bin k in the target image. The L_1 distance between the color histogram of the model and that of the estimated foreground region is

$$H(x,s) = \sum_k |\tilde{h}(k) - s\sum_i h_i(k)x_i|, \qquad (2)$$

where s is the unknown scale. Note that it is critical to apply s to the target histogram. The seemly simpler formulation that applies s to the model histogram has a bias to find small object foreground. By minimizing H, the foreground estimation matches model's global color composition.

H is nonlinear. We convert it into a linear form. We use the trick $\min |x| \Leftrightarrow \min y : -y \leq x \leq y, y \geq 0$ to remove the L_1 norm in H:

$$\min H(x,s) \Leftrightarrow \min \sum_k g_k \quad (3)$$
$$\text{s.t. } -g_k \leq \tilde{h}(k) - s\sum_i h_i(k)x_i \leq g_k, g_k \geq 0.$$

In Eq. (3), the objective function is linearized, but the constraints are still nonlinear because of the quadratic term sx_i.

To linearize the quadratic term, we first consider a special case, in which s is the only scaling factor in the formulation. We notice that since x is binary, sx_i takes value s if $x_i = 1$, and otherwise 0. We therefore introduce a new variable u_i and we enforce that $u_i = s$ when $x_i = 1$, and otherwise $u_i = 0$. This can be achieved by introducing the linear constraint

$$s + L(x_i - 1) \leq u_i \leq s + L(1 - x_i),\ 0 \leq u_i \leq Lx_i, \quad (4)$$

where L is a large positive number. With the above constraint, if x_i takes 1, u_i has to equal s, and if x_i is 0, $u_i = 0$. Therefore the quadratic term can be linearized as $s\sum_i h_i(k)x_i = \sum_i h_i(k)u_i$, where u_i satisfies the constraint in Eq. (4).

In general cases, the formulation may involve scaling factors that are nonlinear functions of s. In section 2.1.2, we introduce the perimeter consistency term that has a scaling factor $s^{0.5}$ instead of s. In this case, the above continuous scale method is not able to generate a complete linear formulation. To solve the problem, we quantize s into discrete values. Instead of labeling a superpixel as foreground or background, we label each superpixel as foreground or background at a specific scale $l_m, m = 1..M$, where l_m is a quantized s at level m. The augmented labeling variable is denoted as $x_{i,m}$, which is 1 if superpixel i is labeled as foreground at scale level m, and otherwise 0. We have $\sum_m x_{i,m} = x_i$, if we collapse $x_{i,m}$ along the axis of scale. x_i and $x_{i,m}$ are correlated by

$$x_{i,m} \leq x_i,\ x_{i,m} \leq s_m,\ x_{i,m} \geq x_i + s_m - 1,\ x_{i,m} \geq 0, \quad (5)$$

where s_m is a binary variable that takes 1 if scale level m is selected and 0 otherwise. Since we can only choose one scale, $\sum_{m=1}^{M} s_m = 1$. Following the above procedure, we convert the quadratic term into a linear expansion: $s\sum_i h_i(k)x_i \approx \sum_{i,m} l_m h_i(k)x_{i,m}$. H is linearized.

2.1.2 Shape Consistency

Apart from color histogram matching, we require that the segmentation is consistent with both the area and the perimeter of the model at scale s. This enforces a shape constraint on the estimated foreground region.

The area consistency term is represented as follows. Let $\tilde{a}$ be the area of the model and a_i be the area of superpixel i. The area consistency term $A(x,s)$ is

$$A(x,s) = |\tilde{a} - s\sum_i a_i x_i| \approx |\tilde{a} - \sum_{i,m} l_m a_i x_{i,m}|, \quad (6)$$

which can be turned into linear objective and constraints using the auxiliary variable trick used in Eq. (3).

We further constrain the perimeter of the foreground estimation. In superpixel labeling, a boundary line between two adjacent superpixels becomes part of the object boundary if the superpixels receive different labels. If a superpixel is located adjacent to the boundary of the image and labeled as foreground, its boundary segment attaching to the image boundary should also be counted. To simplify the formulation, we introduce a dummy region to represent everything outside of the image and fix its label as 0; in this way, the border superpixels can be treated just as other regular ones. We minimize the difference of the perimeter of the target region with the model at scale s. Note that when an object's area is scaled by s, its perimeter is scaled by $s^{1/2}$. The perimeter consistency term P is

$$P(x,s) = |\tilde{b} - s^{1/2} \sum_{\{i,j\}\in\mathcal{N}} b_{i,j}|x_i - x_j|\,|, \quad (7)$$

where $\tilde{b}$ is the model boundary length and $b_{i,j}$ is the boundary length between adjacent superpixels i and j. The set of adjacent superpixels is $\mathcal{N}$. To linearize P, we introduce pairwise variable $y_{i,j}$ that is $s^{0.5}$ if x_i and x_j take different values, and 0 otherwise. y and x are related by

$$y_{i,j} \leq s^{0.5}(x_i + x_j),\ y_{i,j} \leq s^{0.5}(2 - x_i - x_j), \quad (8)$$
$$y_{i,j} \geq s^{0.5}(x_i - x_j),\ y_{i,j} \geq s^{0.5}(x_j - x_i).$$

It is not hard to verify that with the above constraints, if x is binary, y must be the desired pairwise assignment variable. Notice that $s^{0.5} \approx \sum_m l_m^{0.5} s_m$ and $s^{0.5}x_i \approx \sum_m l_m^{0.5} x_{i,m}$, and by substitution we linearize the constraints in Eq. (8). The perimeter term becomes

$$P = |\tilde{b} - \sum_{\{i,j\}\in\mathcal{N}} b_{i,j} y_{i,j}|. \quad (9)$$

Since we use L_1 norm, by further using auxiliary variables, the P term can be finally converted to linear.

2.1.3 Optional Linear Terms

We can further include optional unary and pairwise terms. The unary term quantifies the local similarity of each superpixel to the model, and is written as $\sum_i c_i x_i$, where c_i is sum of the smallest distance from each color in superpixel i to the model colors. The pairwise term quantifies how strongly two superpixels connect and we prefer to merge similar superpixels with weak boundaries. The merging cost is thus determined by color histogram similarity and the contact edge strength. The cost of merging neighboring superpixels i and j is defined as $d_{i,j} = D(i,j) + \kappa g_{i,j}$, where $D(i,j)$ is the χ^2 distance between the color histograms of superpixels i and j, and $g_{i,j}$ is the average gradient magnitude on the boundary between superpixels i and j. In this paper, κ is 1. T is written as

$$T = \sum_{i=1}^{N} c_i x_i + \sum_{\{i,j\}\in\mathcal{N}} d_{i,j} x_i x_j, \quad (10)$$

in which $x_i x_j$ can be replaced by variable $t_{i,j}$ that satisfies $t_{i,j} \leq x_i, t_{i,j} \leq x_j, t_{i,j} \geq 0$ and $t_{i,j} \geq x_i + x_j - 1$. Since T has a bias toward small foreground estimation, it has a small weight 0.01.

Combining the above terms we obtain a linear optimization. However, its solution does not guarantee the connectivity of the foreground estimation. We further introduce connectivity cuts into the optimization.

2.1.4 Connectivity Constraint

Connectivity cuts are linear constraints that are violated by non-connected foreground estimations. There are different ways to specify these cuts [15, 11]. Even though our objective function in Eq. (1) is nonlinear and contains high order terms, these connectivity cuts can still be applied. In this paper, we adopt the linear cut proposed in [11]. The difference is that we directly work on integer solutions instead of floating point relaxations. There is also no concept of the most violated cut and we need a different criterion to insert cuts.

We construct a superpixel graph whose nodes correspond to superpixels. We insert bidirectional edges between two nodes if the corresponding superpixels are adjacent. In the superpixel graph, the set of bottleneck nodes for superpixels i and j, denoted by $Q_{i,j}$, includes all the nodes whose corresponding superpixels have been labeled as background and would disconnect the path from i to j if all of them are removed. If i and j are two disconnected foreground superpixels, the connectivity cut is $x_i + x_j - \sum_{k \in Q_{i,j}} x_k \leq 1$. The bottleneck nodes can be found by a maxflow-mincut algorithm. Different from [11], we build the flow network directly on top of the superpixel graph and set edge capacity to 1 if the corresponding adjacent superpixels have at least one 0 label, and otherwise the edge capacity is infinity. Using the flow network, we compute the maxflow between a pair of nodes that correspond to disconnected foreground superpixels i and j; with the residual network, we find the cut edges that separate the source node cluster and sink node cluster; the bottleneck nodes in $Q_{i,j}$ are the source nodes of these cut edges if they do not correspond to superpixel i or j, or else their target nodes are the bottleneck nodes.

For disconnected foreground superpixels the left hand sides of all the cut inequalities equal 2. We therefore need a strategy to choose which cut to include into the optimization. In this paper, we introduce the cut that corresponds to the top two largest disconnected foreground regions in the current solution. Our experiment shows that this method has the fastest convergence rate comparing to other approaches.

2.2. Optimization

The complete mixed integer linear formulation for the global foreground estimation is:

$$\min\{\sum_k g_k + \lambda w + \mu p + \gamma(\sum_i c_i x_i + \sum_{\{i,j\} \in \mathcal{N}} d_{i,j} t_{i,j})\} \quad (11)$$

Subject to:

$$-g_k \leq \tilde{h}(k) - \sum_{i,m} l_m h_i(k) x_{i,m} \leq g_k,\ g_k \geq 0$$

$$x_{i,m} \leq x_i,\ x_{i,m} \leq s_m,\ x_{i,m} \geq x_i + s_m - 1,\ \sum_m s_m = 1$$

$$-w \leq \tilde{a} - \sum_{i,m} l_m a_i x_{i,m} \leq w,\ w \geq 0$$

$$y_{i,j} \leq \sum_m l_m^{0.5}(x_{i,m} + x_{j,m}),\ y_{i,j} \leq \sum_m l_m^{0.5}(2 - x_{i,m} - x_{j,m}),$$

$$y_{i,j} \geq \sum_m l_m^{0.5}(x_{i,m} - x_{j,m}),\ y_{i,j} \geq \sum_m l_m^{0.5}(x_{j,m} - x_{i,m}),$$

$$-p \leq \tilde{b} - \sum_{\{i,j\} \in \mathcal{N}} b_{i,j} y_{i,j} \leq p,\ p \geq 0,$$

$$x_i \geq t_{i,j},\ x_j \geq t_{i,j},\ t_{i,j} \geq 0,\ t_{i,j} \geq x_i + x_j - 1,\ \forall \{i,j\} \in \mathcal{N}$$

Connectivity Cuts, $x_i, s_m = 0$ or 1, all variables ≥ 0

The linear optimization is equivalent to the original nonlinear optimization in Eq. (1) on discrete scales. Eq. (11) has the following structure:

$$\min e^T z : Fz = f, Dz = d \quad (12)$$

where z is a vector that includes x, y, s and other auxiliary variables and e is the constant coefficient vector. The fixed constraint $Fz = f$ is induced by the color histogram, area, perimeter, unary and pairwise terms, and the dynamic constraint $Dz = d$ is composed of the connectivity cuts. The set of connectivity cuts starts from an empty set and expands with a single cut at a time to penalize the two largest disconnected foreground regions. Therefore the proposed method finds object foreground by iteratively solving a sequence of mixed integer linear programs. This procedure terminates when there is only one connected foreground region in the estimation.

Proposition 1. The above iterative cutting plane procedure guarantees that each connected foreground is a feasible solution of the integer linear program. And, as the iteration terminates, we obtain the optimal connected segmentation that minimizes the objective in Eq. (1) on discrete scales.

We never lose feasible connected solutions when we shrink the feasible region of Eq. (11) by adding connectivity cuts, since each connected foreground estimation has to satisfy every new cut introduced. Therefore the optimum of the integer program is not greater than the global optimum of Eq. (1) on discrete scales. Since there are finite number of connectivity cuts, the iteration terminates in finite number of steps. As the iteration terminates, the result is a connected foreground estimation and therefore its cost is not smaller than the global optimum of Eq. (1) on discrete scales. The sequential mixed integer program is thus equivalent to solving the original nonlinear optimization on discrete scales.

We use a branch and bound procedure to solve the sequence of mixed integer linear programs. The key for fast

convergence is to find tight upper and lower bound quickly. The lower bound of the mixed integer linear program is obtained by its relaxation that ignores the integer constraint. If the linear program gives integer solution for x_i and s_m, the objective is an upper bound. Otherwise, we find an upper bound by rounding. For optimization without connectivity cuts, the upper bound is found by rounding the linear program solution for $x_i, i = 1..N$, with a threshold of 0.5 and searching through each scale assignment for s_m to minimize the objective. With connectivity cuts, we obtain a binary solution that satisfies all the constraints by thresholding x_i with 0.5 and finding the largest connected foreground component on the superpixel graph as the object foreground estimation; we then try each discrete scale and pick up the smallest objective as an upper bound estimation. We update the upper bound if the current estimation is smaller.

We branch the search tree on the node with the smallest lower bound and introduce integer cuts on the most fractional variable (the variable closest to 0.5). For the node with the lowest lower bound, a new cut $z_i = 0$ or $z_i = 1$ where z_i is either an x variable or an s variable is included in the linear program. Fortunately, we do not have to solve each linear program from scratch, since there is only one more new constraint included. By introducing slack variables, $z_i = 0$ or equivalent $z_i \leq 0$ becomes $z_i + v_{i,0} = 0$, and $z_i = 1$ or equivalent $z_i \geq 1$ becomes $z_i - v_{i,1} = 1$ where $v_{i,0} \geq 0, v_{i,1} \geq 0$. z_i is a basic variable and its right hand side is a fractional number in the final simplex tabular. For the $z_i = 0$ branch, we subtract the original z_i row from $z_i + v_{i,0} = 0$, and for the $z_i = 1$ branch, we subtract $z_i - v_{i,1} = 1$ from the z_i row. In either case, we turn $v_{i,0}$ or $v_{i,1}$ into a basic variable that is non-feasible because it has negative value on the right hand side. The dual-simplex method is then applied in pivoting and usually it takes very few steps to regain the optimal solution.

We discard the branch whose linear program solution is infeasible or is greater than the current upper bound. Most of the branches are pruned quickly. We keep track of the upper bound B_u and lower bound B_l of the solution. B_l is the lowest lower bound of all the active search tree nodes. Branch and bound can be terminated prematurely when the tolerance gap $\delta = \frac{2(B_u - B_l)}{(B_u + B_l)}$ is reached. Fortunately, the quality of segmentation degrades slowly as δ increases. We therefore can choose large δ to speed up the search with only minor performance degradation. The branch and bound procedure essentially solves a set of promising linear relaxations with different choices of fixed labels and finally we pick the best result. When the search is complete, the objective is upper bounded by $(\delta + 2)/(2 - \delta)$ times the global minimum. To further reduce the complexity, we can also terminate the search when the foreground connectivity ratio is above a threshold r. We define the connectivity ratio as the percentage of the area of the largest connected component in the whole foreground area. The branch and cut procedure is summarized as follows.

Algorithm 1. Linear Global Figure Ground Separation

1. Compute superpixels, their color histograms, areas, perimeters, local costs and merging costs. Set tolerance gap δ and connectivity ratio r.
2. Construct the mixed integer linear program (Eq. 11).
3. Branch and bound with gap tolerance δ (section 2.2).
4. While *foreground connectivity ratio* $< r$

   Include a connectivity cut (section 2.1.4).

   Branch and bound with δ (section 2.2).
5. Find the largest connected foreground component and let it be the estimated object foreground.

3. Experimental Results

We first illustrate the advantage of the proposed scale invariant approach over explicitly searching through multiple scales. Fig. 2 compares these two approaches. In our experiment, we use the graph method in [16] to generate superpixels. In multiple scale segmentation, we try each quantized scale explicitly and apply the branch and cut method in each case. Fig. 2 rows 2-5 show results of the multiple scale method. Fig. 2 row 6 shows the result of the proposed scale invariant method with the same parameter setting ($\delta = 0.5, r = 1$). We use 8 bins for each color channel. The proposed method converges with three connectivity cuts to satisfy the connectivity constraint and there are total 380 linear programming relaxations in the branch and bound comparing to total 4452 linear relaxations in the multiple scale method. The scale invariant method is 10 times faster than the multiple scale approach.

3.1. Synthetic Data

We test the proposed method on synthetic ground truth data. The synthetic images contain blocks with random intensities. The foreground is generated with a random mask that roughly has a rectangle shape. Fig. 3 (a) shows a foreground mask. Fig. 3 (b) is an image randomly generated using the mask. The foreground intensity is uniformly distributed in $[0, 1]$. Clutter with similar intensities to the foreground is introduced into the background. We include random perturbation to the foreground intensities to simulate object appearance variations. The template is randomly scaled in $[0.5, 2]$. We use two levels of foreground color disturbance 0.01 and 0.02, and 3 levels of clutter settings 0.1, 0.15 and 0.25. Therefore there are total 6 test cases. For each test case, we randomly generate 1000 test images. We compare the proposed method with its variants and several competing method.

We first test how the settings on the tolerance gap δ would affect the segmentation result. Fig. 3 (c) and (d) show foreground estimations of the sample image in Fig. 3 (b) using $\delta = 10^{-6}$ and 1. We obtain the same segmentation result under these two settings, while the second one is many times faster. In Fig. 4, the bar charts for ThisPaper (a), (b)

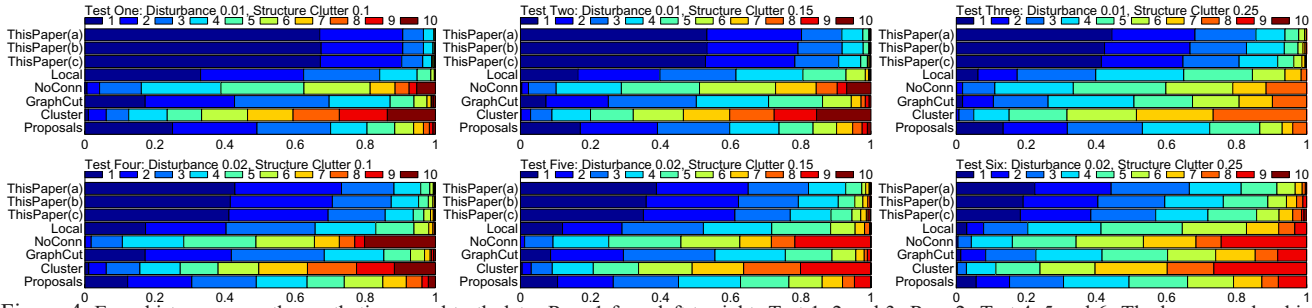

Figure 4. Error histograms on the synthetic ground truth data. Row 1 from left to right: Test 1, 2 and 3. Row 2: Test 4, 5 and 6. The lower number bins correspond to lower errors. Bar length represents the proportion in all the tests. Good results correspond to longer bars for low-number bins and shorter bars for high-number bins. ThisPaper (a), (b) and (c) correspond to the proposed method with $\delta = 10^{-6}, 0.5$ and 1 respectively. Local method uses local feature similarity and unary, pairwise terms, and maintains foreground connectivity. NoConn ignores the connectivity constraint. We also compare with GraphCut [17], Cluster [8] and Proposals [5] methods.

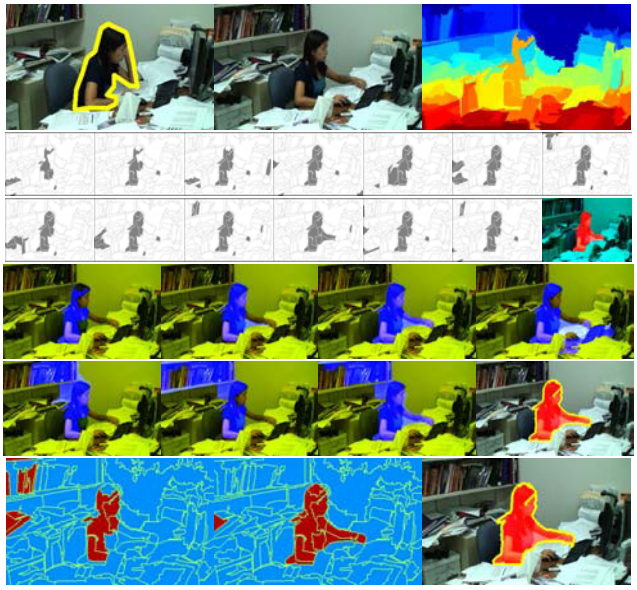

Figure 2. Comparison of scale invariant method with multiple scale method. Row 1: Model object, target image and superpixels. Row 2-3: Samples of 17 iterations to connect foreground estimation at scale 1. Row 4-5: Foreground estimations for scales from 0.5 to 2 (blue) and the final result (red). It takes about 10 second to complete. Row 6: The proposed scale invariant method satisfies the foreground connectivity constraint in 3 iterations and converges in less than 1 second. It uses total 380 relaxations comparing to 4452 in the multiple scale approach.

and (c) show the error distributions of the proposed method with $\delta = 10^{-6}, 0.5$ and 1 in 6000 tests. The error metric is defined as follows. We compute the sum of the absolute values of the pixel-wise differences between the foreground estimation map and the ground truth map, and use its ratio to the total number of the ground truth foreground pixels as the error measurement. The ratio may be greater than 1. As shown in Fig. 4 and Fig. 5, even with a large error bound setting, the segmentation result degrades little. By setting large δ, the running time reduces by orders of magnitude.

We compare with several competing methods. The first competing method uses local feature similarity in the objective function with unary and pairwise terms. It guarantees the foreground connectivity. The second competing method is the variant of the proposed method that does not include

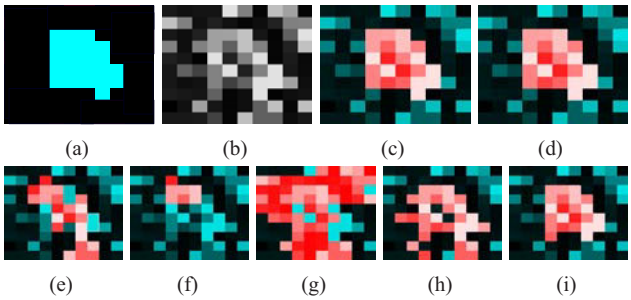

Figure 3. Sample results on synthetic ground truth data. (a) A foreground mask. (b) A random pattern generated from (a). (c) The proposed method with $\delta = 10^{-6}$ completes in 1.9 seconds, and (d) with $\delta = 1$ it completes in 0.09 seconds with the same result. The second row shows the results of competing methods: (e) local method, (f) global method ignoring connectivity constraint, (g) the graph-cut based method [17], (h) cluster method [8], and (i) object category independent proposal method [5].

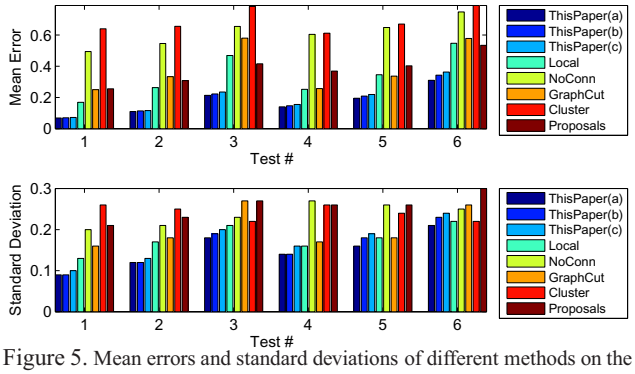

Figure 5. Mean errors and standard deviations of different methods on the synthetic ground truth data.

the connectivity constraint. If the foreground estimation is not connected, we find connected component that has the lowest objective in Eq.(1). We compare the proposed method with the iterative graph-cut method [17], discriminative clustering method [8] and object category independent proposal method [5]. The clustering method does not consistently set the foreground to 1, so we try both the 1 region and 0 region and find the best estimation. If the cluster and graph cuts based methods generate disconnected foreground estimations, we enumerate all the connected components using the criterion the same as the proposed method in Eq. (1) except the superpixel merging term. Using the

Figure 6. Sample results of the proposed method on 97-frame waterski sequence [7]. Upper-left image shows the template.

Figure 7. Tracking using video snapcut [14]. Samples results are for bike sequence (row 1) and TV show sequence (row 2).

method in [5] we generate proposals of foreground segmentation and then find the best one using Eq. (1) without the superpixel merging term. We adjust color histogram bin numbers for the competing methods to achieve their best performance. Competing methods also have the advantage of choosing non-connected foreground estimation if it matches the ground truth better than their connected estimation. Sample comparison results on the synthetic data are shown in Fig. 3. Fig. 4 and Fig. 5 show the quantitative comparison results; the proposed method has the lowest mean errors in all the tests.

3.2. Real Images

We test the proposed method on challenging real images and videos. Fig. 6 shows sample results of the proposed method on the waterski sequence from [7]. The model is generated from a randomly selected image. The proposed method successfully segments the object even though there are large scale changes, occlusion and motion blur. In the real image experiments, we set $\delta = 0.5, r = 1$ and use 8 bins in each color channel.

For quantitative comparison, we test on 7 challenging video sequences (Fig. 10) including the figure skating sequence from [1], skating and dance videos from [13], a baby video from YouTube, two broadcasting videos and a recorded video. These videos include strong clutter, large scaling, object deformation, and some contain abrupt camera view angle changes. Traditional video segmentation methods that reply on motion continuity have difficulty in dealing with such sequences. Fig. 7 shows sample results of the video snapcut [14] (Roto Brush in Adobe After Effects) on the bike and TV show sequence with manual initialization in the first video frame. Video snapcut drifts and gradually loses the target in the bike sequence, and it is not able to segment the object in the TV show sequence due to abrupt video cuts. In contrast, the proposed method is able to give reliable results as shown in Fig. 10 row 4 and row 6.

We compute the object foreground model from a randomly selected image in each test sequence. The model images and the labeled foregrounds are shown in Fig. 8

Figure 8. Sample results on 7 test videos. Foreground is shown in red channel. Row 1: Model images and objects of interest. Row 2: Proposed method. Row 3: Ignoring connectivity constraint. Row 4: Local method. Row 5: Cluster method [8]. Row 6: Graph-cut method [17]. Row 7: Proposal based method [5].

row one. Fig. 8 shows how the proposed method improves the results over the five competing methods. The competing methods use the same settings as those in section 3.1. For quantitative comparison, we labeled half of the video frames in each sequence and we measure the mean errors and the error standard deviations of each competing method. The error metric is the same as the one that we use in the synthetic data experiment. Fig. 9 summaries the mean errors and standard deviations of different methods on the real image test. The proposed method consistently gives the smallest mean errors in all the tests. Overall, it also has the smallest standard deviation.

More sample results of the proposed method on the test sequences are shown in Fig. 10. The proposed method is able to handle strong clutter, scale changes, motion blur, large deformation and object view angle changes. Since color histogram and the shape features are resistant to object deformation, and different terms are combined in a soft fashion in the objective function, our method is not sensitive to the model image selection. Local segmentation errors are mostly caused by strong background clutter that has similar appearance to the model and is connected to the object foreground. The result also degrades if the superpixels are too coarse. Fortunately, the proposed method allows us to use detailed over-segmented superpixels and still maintains efficiency. The typical running time of the proposed method on each image is less than 5 seconds on a 2.8GHz machine, which is similar to the graph-cut method [17] and the local approach, and is faster than the proposals [5] and clustering [8] methods that take about 200 seconds per image.

4. Conclusion

We propose a novel scale invariant linear solution to global figure ground separation. The proposed method explicitly enforces the global color histogram, area, perimeter and connectivity constraints. It requires only the foreground model and therefore is able to segment objects in changing backgrounds. The solution of the proposed method has a guaranteed error bound. Our experiments on both synthetic

Figure 10. Sample results of the proposed method on 7 test sequences. Row 1: 751-frame skate-I. Row 2: 493-frame skate-II. Row 3: 651-frame baby. Row 4: 501-frame bike. Row 5: 1016-frame table. Row 6: 600-frame TV show. Row 7: 144-frame dance.

	Skate-I	Skate-II	Baby	Bike	Table	TVShow	Dance
This Paper	**0.338 ±0.09**	**0.309 ±0.09**	**0.314 ±0.08**	**0.442** ±0.18	**0.264** ±0.09	**0.225 ±0.12**	**0.271 ±0.15**
NoConn	0.447 ±0.16	0.516 ±0.44	0.538 ±0.16	0.738 ±0.56	0.434 ±0.33	0.475 ±0.42	0.361 ±0.24
Local	2.122 ±1.07	3.000 ±4.37	0.912 ±0.30	4.283 ±4.21	1.690 ±0.64	2.303 ±1.52	2.588 ±0.93
Cluster [8]	2.045 ±2.14	1.020 ±0.26	1.222 ±0.44	1.039 **±0.15**	1.000 **±0.01**	0.965 ±0.14	3.781 ±0.99
GraphCut [17]	0.383 ±0.15	0.872 ±1.43	0.979 ±0.28	0.946 ±0.79	0.914 ±0.27	0.608 ±0.47	2.376 ±0.93
Proposal [5]	0.512 ±0.19	0.334 ±0.17	1.129 ±0.68	0.687 ±0.39	0.675 ±1.05	0.599 ±0.28	2.162 ±2.21

Figure 9. Mean errors and standard deviations of different methods on 7 test videos. Bold fonts indicate the smallest value in each category.

ground truth data and challenging real videos show that the proposed method is efficient and has superior performance to the competing methods.

Acknowledgment: This research is supported by US NSF grant 1018641.

References

[1] X. Ren and J. Malik. Tracking as Repeated Figure/Ground Segmentation. CVPR 2007. 1, 7

[2] D. Tsai, M. Flagg, and J.M. Rehg. Motion Coherent Tracking with Multi-Label MRF Optimization. BMVC 2010. 1

[3] J. C. Niebles, B. Han and L. Fei-Fei. Efficient Extraction of Human Motion Volumes by Tracking. CVPR 2010. 1

[4] S. Vijayanarasimhan, K. Grauman. Efficient Region Search for Object Detection. CVPR 2011. 1

[5] I. Endres and D. Hoiem. Category Independent Object Proposals. ECCV 2010. 1, 6, 7, 8

[6] Y.J. Lee, J. Kim, K. Grauman. Key-Segments for Video Object Segmentation. ICCV 2011. 1

[7] M. Grundmann, V. Kwatra, M. Han, I. Essa. Efficient Hierarchical Graph-Based Video Segmentation. CVPR 2010. 1, 7

[8] A. Joulin, F. Bach, J. Ponce. Discriminative Clustering for Image Co-segmentation. CVPR 2010. 2, 6, 7, 8

[9] C. Rother, T. Minka, A. Blake, V. Kolmogorov. Cosegmentation of Image Pairs by Histogram Matching - Incorporating a Global Constraint into MRFs. CVPR 2006. 2

[10] L. Mukherjee, V. Singh, C.R. Dyer. Half-integrality Based Algorithms for Cosegmentation of Images. CVPR 2009. 2

[11] S. Nowozin, C. Lampert. Global Connectivity Potentials for Random Field Models. CVPR 2009. 1, 4

[12] S. Todorovic, N. Ahuja. Unsupervised Category Modeling, Recognition, and Segmentation in Images. TPAMI, v.30, n.12, 2008. 2

[13] H. Jiang, T. Tian and S. Sclaroff. Scale and Rotation Invariant Matching Using Linearly Augmented Trees. CVPR 2011. 2, 7

[14] X. Bai, J. Wang, D. Simons, G. Saprio. Video SnapCut: Robust Video Object Cutout Using Localized Classifiers. SIGGRAPH 2009. 7

[15] I. Ljubic, R. Weiskircher, U. Pferschy, G. Klau, P. Mutzel, M. Fischetti. An Algorithmic Framework for the Exact Solution of the Prize-collecting Steiner Tree Problem. Math. Prog. 2006. 1, 4

[16] P. Felzenszwalb, D. Huttenlocher. Efficient Graph-Based Image Segmentation IJCV, Vol. 59, No. 2, 2004. 5

[17] I.B. Ayed, H. Chen, K. Punithakumar, I. Ross, S. Li. Graph Cut Segmentation with a Global Constraint: Recovering Region Distribution via a Bound of the Bhattacharyya Measure. CVPR 2010. 2, 6, 7, 8