Multiple Foreground Recognition and Cosegmentation: An Object-Oriented CRF Model with Robust Higher-Order Potentials

Hongyuan Zhu[1]*, Jiangbo Lu[2], Jianfei Cai[1], Jianming Zheng[1], and Nadia M. Thalmann[1]

[1]Nanyang Technological University, Singapore
[2]Advanced Digital Sciences Center, Singapore

Abstract

Localizing, recognizing, and segmenting multiple foreground objects jointly from a general user's photo stream that records a specific event is an important task with many useful applications. As argued in recent Multiple Foreground Cosegmentation (MFC) work by Kim and Xing, this task is very challenging in that it contrasts substantially from the classical cosegmentation problem, and aims to parse a set of realistic event photos but each containing irregularly occurring multiple foregrounds with high appearance and scene configuration variations. Inspired by the impressive advance in scene understanding and object recognition, this paper casts the multiple foreground recognition and cosegmentation (MFRC) problem within a conditional random fields (CRFs) framework in a principled manner. We capitalize centrally on the key objective that MFRC is to segment out and annotate foreground objects or "things" rather than "stuff". To this end, we exploit a few complementary objectness cues (e.g. contours, object detectors and layout) and propose novel and efficient methods to capture object-level information. Integrating object potentials as soft constraints (e.g. robust higher-order potentials defined over detected object regions) with low-level unary and pairwise terms holistically, we solve the MFRC task with a probabilistic CRF model. The inference for such a CRF model is performed efficiently with graph cut based move making algorithms. With a minimal amount of user annotations on just a few example photos, the proposed approach produces spatially coherent, boundary-aligned segmentation results with correct and consistent object labeling. Experiments on the FlickrMFC dataset justify that our method achieves state-of-the-art performance.

1. Introduction

With the popularity of digital cameras and mobile phones, it becomes very easy now for people to record their daily life in a visually rich way. How to effectively manage, understand and exploit a set of photos a user takes for a certain event is a very interesting topic, leading to many exciting applications. Typically, such event photos contain multiple foreground objects of interest, but only an unknown number of these objects appear irregularly in each photo, where the background may also vary. This paper concerns the problem of localizing, recognizing, and segmenting multiple foreground objects jointly from a general user's photo stream. We refer to such a problem as *Multiple Foreground Object Recognition and Cosegmentation* (MFRC). This MFRC task is challenging due to strong intra- and inter-object variation, background clutter and sharing features among different classes of objects, to name a few, in addition to the irregular foreground occurrence patterns mentioned earlier. Our work is motivated by the recent study by Kim and Xing [15], but it significantly advances the MFRC performance with several novel techniques.

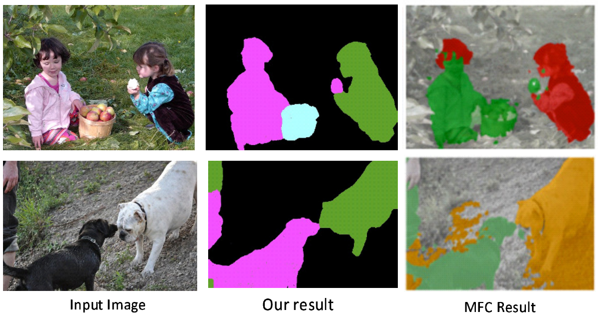

Figure 1. Visual comparison between the state-of-the-art MFC method [15] and our method on two images from the FlickrMFC dataset. The MFC method wrongly annotates the *apple+bucket* foreground as the *girl+red*, and the hair of the *girl+red* foreground as the *girl+blue* foreground. It also misclassifies the background as the *dog+white* in the second test image. Our method does a much better job in resolving the ambiguity in the MFRC task with a hierarchical CRF model using higher-level object cues.

While many cosegmentation algorithms [15, 16, 13, 29, 14, 8, 26, 25] exist, most of them are built upon the assumption that the same objects appear in all input photos, which is easily violated in a general MFRC scenario. Recently, Kim and Xing [15] proposed an approach specifically designed to address this MFRC problem, and achieved superior results in comparison with other existing methods. However, though making the solution tractable, their design counting on coarse segmentation and the restriction imposed on generating foreground candidates are often

*This work was mainly done when Hongyuan was an intern at ADSC.

over-simplified treatments and give only a sub-optimal solution for complicated realistic scenes. As a result, this method cannot generate accurate recognition and segmentation results consistently for more challenging MFRC cases, especially when more high-level, non-local interactions are needed to resolve the ambiguity (see Fig. 1).

Inspired by the impressive recent advance in scene understanding [32, 21, 18], object recognition, detection and segmentation [11, 10, 2, 7, 20, 19], we cast the MFRC problem similarly within a conditional random fields (CRFs) framework in a principled manner. At the heart of our proposed approach is the integration of the objectness notion into a probabilistic CRF model. Our key observation is that in general the common goal of MFRC is to segment out and annotate foreground objects or "things" (e.g. girl dressed in red, apple bucket) rather than "stuff" (e.g. sky, grass). Similar ideas of incorporating object-like proposals [2] or object detectors [11, 10] in a conventional CRF framework have been successfully applied before to other vision tasks such as large-scale image segmentation [20] and scene understanding [22]. However, the MFRC task considered here is unique and very challenging – the user only gives a minimal amount of annotations on just a few example photos, while the possible geometric and photometric variations that irregularly occurring multiple foregrounds exhibit across the photo set can be quite large. This paper is hence triggered to answer how far we can achieve for the challenging yet useful MFRC task, leveraging recent advances from object detection [35] to robust higher-order CRFs inference [18].

In this paper, we propose a few robust and complementary objectness cues and object-based labeling consistency constraints (e.g. contours, multi-class object detectors, layout patterns), and combine them with low-level unary and pairwise terms holistically in a CRF model. We further augment the CRFs by including robust higher-order potentials defined over detected object regions, which is beneficial to inference results but also can be solved efficiently with graph cut based move making algorithms [6]. Experiments on the FlickrMFC dataset demonstrates state-of-the-art performance of the proposed algorithm, which generates spatially coherent, boundary-accurate segmentation results with correct and consistent multiple foreground recognition.

1.1. Relations and Comparison with Previous Work

Cosegmentation. There is a vast amount of prior work on cosegmentation [15, 16, 13, 29, 14, 8, 26, 5, 9]. Most of the existing works focus on handling the binary cases, separating foreground(s) from the background, but few of them are designed for joint multi-class object recognition and segmentation. The unsupervised methods such as DC [14], Cosand [16] and MC [13] used low-level bottom-up features, so they cannot distinguish "stuff" from "objects" in presence of background clutter and sharing features among classes. To overcome the ill-defined nature of unsupervised methods, some user inputs are hence desired and also often necessary. One notable work is iCoseg [5], which solves binary foreground cosegmentation using graph cut. Our proposed algorithm involves a minimal amount of user annotations on a very small fraction of the image set in the form of bounding boxes (or polylines) and object labels. But unlike the aforementioned methods, we do not require the user to carefully sort out a given event photo set manually to group images containing the same objects together. The MFC method [15] is one of the existing works which attempt to solve the irregularly occurring multiple foregrounds problem. Similar with MFC [15], our method also deals with the multiple foreground cosegmentation problem. However, we perform joint detection and segmentation of multiple objects for a set of event photos, which often exhibit high variability of foreground objects in shape, color and their complicated interactions with other objects or varying backgrounds. Technically, our algorithm incorporates the higher level non-local object cues into a probabilistic inference and optimization framework, which has not been explored before in previous cosegmentation works. Such non-local cues, which help to differentiate "stuff" and "objects", are expressed as soft constraints. Thanks to the soft constraint, multiple hypotheses can compete to make our method robust to false positive detection hypotheses, so they do not affect the final results when strongly defended by the hypotheses based on pixels and segments. Recently, Ma and Latecki [23] proposed a semi-supervised graph based method to perform the MFC task with a new connectivity constraint and achieved state-of-the-art on the subset of FlickrMFC dataset. In fact, our work is theoretically complementary to Ma and Latecki's work, which proves that higher order constraints are beneficial for the MFRC task. In addition, our method is scalable to large image datasets, while the method in [23] does not scale well due to its reliance on dense pairwise image analysis. In terms of the experimental results, [23] does not report on the challenging "thinker+Rodin" group existing in the full FlickrMFC dataset, which features challenges such as strong intra object variation, background clutter and lighting and scale changes. In contrast, we reported the results on the full FlickrMFC dataset, and achieved much better accuracy than the MFC method [15] on the "thinker+Rodin" group even by 50%.

Object recognition, localization and segmentation. The last few years have seen impressive progress for several areas such as object recognition [11, 10], generic object localization [2, 7], and object segmentation [20, 19]. For instance, objectness window [2] have been successfully applied to single foreground segmentation propagation in ImageNet [20]. Multiple foreground proposals [7] have also been applied to Sarah *et al.*'s binary foreground cosegmentation work [34] and achieved state-of-the-art results. At the same time, combined multi-class object segmentation and recognition techniques have also been proposed to address the grand challenge of complete scene understanding [32, 22]. Lubor *et al.* [22] proposed to incorporate object detector-induced potentials into a CRF energy optimization framework as a soft constraint, which clearly improved the standard object class segmentation models that tend to underperform on the "things" classes for complex scenes. Inspired by these nice existing techniques, our work, however, also differs from them in several aspects. First of all, as explained earlier, the MFRC task is very unique and challenging due to the high variability of foreground objects across the given set of photos and the minimal supervision that is available. Second, geared towards this MFRC task, our algorithm has integrated and extended some selected technical modules. For example, we used discriminative color features [35] to train multiple object detectors. In addition, contour as an important object-oriented property has been novelly exploited in this paper, which proves its effectiveness in the MFRC task.

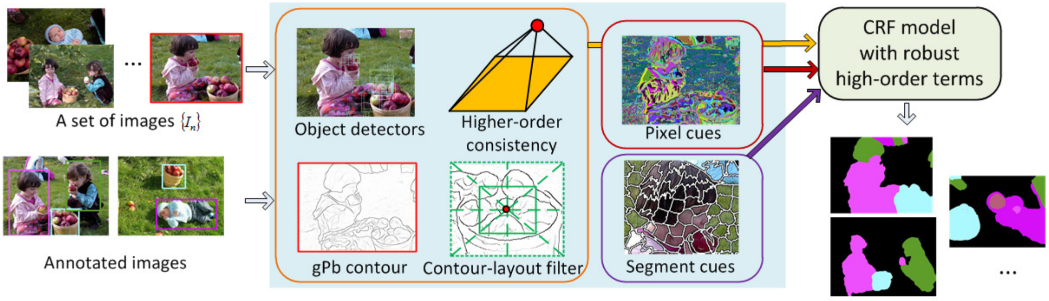

Figure 2. Overview of our MFRC system. It integrates various object-level cues with low-level cues in a probabilistic CRF model.

2. Problem Formulation and Our CRF Model

Given a set of N input images $\mathcal{I} = \{I_1, ..., I_N\}$, $m(m \ll N)$ of them $\mathcal{I}_t = \{I_t^1, ..., I_t^m\} \subset \mathcal{I}$ are first annotated to specify the objects of interest and also their rough spatial extent in the form of bounding boxes or polylines. More specifically, each image from this small training set $\mathcal{I}_t$ with user supervisions contains a subset of annotated objects belonging to K different foregrounds $\mathcal{F} = \{F^1, ..., F^K\}$. Each foreground F^l is associated with a numeric label $l \in \mathcal{L} = \{0, 1, ..., K\}$, where 0 is used to denote the background for notational simplicity. We formulate the MFRC problem in terms of a global energy function defined on a conditional random field (CRF), for which the goal is to assign a random variable x_i to each pixel i in each image a label from $\mathcal{L}$. Our framework integrates various complementary object cues computed from different classifiers learned with low-level features, mid-level edge detectors and an interactive offline bounding box object detector. In fact, the proposed framework also allows to choose any state-of-the-art multi-class object detectors and classifiers, though we will present concrete modules in this paper.

Fig. 2 illustrates the proposed framework, which consists of a few stages and several modules. During the preprocessing stage, various foreground cues such as unary multi-class pixel and segment classifiers, object detectors and gPb contour [4] are modeled and generated. Pixel, segment and object detectors classifiers are trained with user-drawn bounding boxes. The gPb contour map is generated by combining the edge signal from eigenvectors with low-level cues, and it captures mid-level object contours. After the preprocessing stage, based on the gPb signal, we specially design a contour-layout filter to reject false positive detector responses which are very likely to be "stuff". With all the cues computed, we integrate them into a global energy function which enforces the labeling consistency between various level cues and finally produce the solution with fast expansion/move solvers. Once the initial segmentation is generated, our framework supports iteratively updating the learned models and performing the recognition and segmentation tasks to further improve the results.

2.1. Proposed CRFs Framework for MFRC

To make our algorithm linearly scalable with the image set size N, the recognition and segmentation inference is performed individually for each image $I_n \in \mathcal{I}$, similar to MFC [15]. We formulate the MFRC task as a multi-labeling problem with a CRF framework on a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is the set of all image pixels of image I_n, while $\mathcal{E}$ corresponds to the set of all edges defined by a four or eight neighbor system. The proposed probabilistic CRF model is given by a Gibbs energy function as follows:

$$\begin{aligned}\mathbf{E}(\mathbf{x}) = &\sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j) \\ &+ \sum_{s \in S} \psi_s(\mathbf{x}_s) + \sum_{d \in D} \psi_d(\mathbf{x}_d)\ .\end{aligned} \tag{1}$$

In (1), $\mathbf{x}$ denotes any valid label assignment to the random variables $\{x_i\}$, which takes a value from the object label set $\mathcal{L}$. S denotes a superpixel decomposition of the image I_n into a set of disjoint segments $\{s\}$, and $\mathbf{x}_s$ is the clique of pixels covered by the segment s. We denote the set of object detections with D, which are typically returned in the form of bounding boxes enclosing objects. The pixels covered within the d-th detection bounding box are represented as $\mathbf{x}_d$. Our energy function consists of four terms: **1)** the pixel-based unary potential $\psi_i(x_i)$, evaluating the likelihood of a certain label assignment to pixel i; **2)** the pairwise smoothness potential $\psi_{ij}(x_i, x_j)$, penalizing differently labeled adjacent pixels of similar appearance; **3)** the segment-level robust label consistency potential $\psi_s(\mathbf{x}_s)$, charging the label inconsistency cost robustly with the number of variables in the segment s not taking the segment label; **4)** the object detector potential $\psi_d(\mathbf{x}_d)$, enforcing a robust region label consistency constraint that is defined in a similar way to $\psi_s(\mathbf{x}_s)$. These terms collectively capture the information for image/object representation and understanding from different levels in a complementary way. We will elaborate the last two terms modeled as robust high-order potentials in Sect. 3 and 4. A contour-based pairwise smoothness potential $\psi_{ij}(x_i, x_j)$ that improves the standard contrast-sensitive implementation [28] will be presented in Sect. 3.3.

Pixel-based unary potential. The first term $\psi_i(x_i)$ is a unary potential defined on each pixel which indicates its cost of being assigned to a label $l \in \mathcal{L}$:

$$\psi_i(x_i) = -\omega_{pix} \log P(x_i | \mathcal{C}_{pix}); \tag{2}$$

where ω_{pix} is the weighting factor. $P(x_i | \mathcal{C}_{pix})$ denotes a normalized distribution returned by a random forest classifier $\mathcal{C}_{pix}$, which is an ensemble of weak decision trees [3]. The classifier is trained with the pixel-level features whose corresponding labels provided by the user. The features defined on each pixel is a seven-dimensional vector, which consists of six color features (RGB and L^*ab) and one texton feature. We generate textons by convolving the image with 17-dimensional filter banks at different scales, as in

[32]. Then the filter bank responses are clustered using K-means algorithm into T_c code words to generate a texton map which encodes the final pixel-wise texton feature.

3. Incorporating Object Cues

This section presents a few complimentary object cues extracted with different technology, which characterize different aspects of an object in the proposed CRF model for the MFRC task. We also discuss the methods to define the corresponding object-oriented potentials.

3.1. Fast Object Detectors with Boosted Color Bins

The appearance of an image patch/segment by itself is often ambiguous when different objects and background contain similar local features, as it is incapable to capture the global configuration information of object class instances. This motivates us to address the MFRC challenge from higher and longer range grouping levels which have been proved to be useful in some image summarization and scene understanding research [18, 21]. A popular approach is to reason about the objects of interest with the help from rectangular bounding boxes which are generated from some detection methods [11, 33, 10]. But, such detections usually require a large number of training examples and often pose strong structured spatial layout constraints. Though deformable part models [10] can relax the rigid spatial configuration constraint, they are typically slow and not suitable for the MFRC task which shall parse a comparatively small set of images but with strong object variations.

To obtain bounding box proposals more robustly with the invariance to e.g. scale, rotation and non-rigid motion, we train a multi-class interactive offline color based object detector. Given a user drawn bounding box, we adopt the method of Wei *et al*. [35] by projecting all pixel colors onto a set of one dimensional (1D) lines in the RGB color space. These lines have different directions and pass through the point (128, 128, 128). The directions in color space are evenly sampled by 13 lines and then a 1D (normalized) histogram of the projected values is calculated on each line. We also use eight bins for each histogram through an empirical comparison and treat all $13 \times 8 = 104$ color bins as our features. Such features can be extracted using integral histogram [27] very efficiently in a constant time. For the multiple foreground recognition problem considered here, we use the Joint Boosting algorithm rather than Adaboosting adopted in [35], and train a multi-class bounding box classifier. The details of our learning procedure resembles closely with those described in [32]. Similar to [35], our training examples are generated from the user annotated images, which however have multiple class labels. To generate more positive samples and also be robust to object variations across images, the same appearance perturbation scheme [35] is employed, which perturbs the position of the object rectangles randomly by a small amount. Our negative examples are randomly sampled around the non-selected foreground regions using the bounding boxes of the same size as the user-specified ones. The bounding box proposals are generated by sweeping the object windows for a set of predefined scale levels in a test image. They are evaluated by the trained multi-class classifiers, whereas only the top-scored bounding boxes are retained.

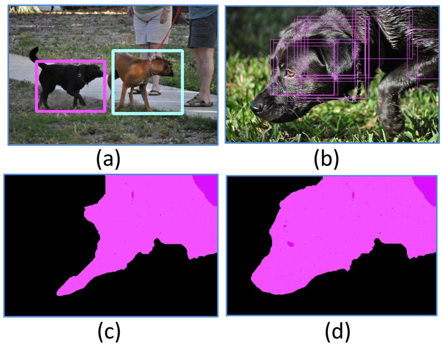

Figure 3. Effects of the object detector-based robust consistency potential. (a) A user-annotated image with two foreground instances labeled with bounding boxes. (b) Applying a learned object detector to a novel image. Shown are top-scored bounding boxes. Segmentation result (c) without and (d) with using the proposed object detector-based label consistency potential.

3.2. Detector-Based Robust Consistency Potentials

A big difference between our energy function and that of conventional binary foreground segmentation is the higher-order bounding-box level potential involved. With the higher-order object information from bounding boxes, we can revolve some ambiguity which would otherwise be too hard to solve at a local level. The bounding box proposals are used to define a kind of soft constraint which works jointly with other hypotheses. We incorporate the object potential $\psi_d(\mathbf{x}_d)$ into our CRF framework by enforcing it as a robust region label consistency constraint defined in [18]. Given the d-th detection bounding box $\mathbf{x}_d$ with a score R_d and the detected object label l_d, $\psi_d(\mathbf{x}_d)$ is defined as:

$$\psi_d(\mathbf{x}_d) = \begin{cases} N_d \frac{1}{Q_d} \gamma_{max} & \text{if } N_d \leq Q_d \\ \gamma_{max} & \text{otherwise} \ , \end{cases} \tag{3}$$

where $N_d = \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)$ is the number of variables in $\mathbf{x}_d$ not taking the dominant label l_d. The truncation parameter Q_d controls the maximum number of inconsistent pixels. The cost γ_{max}, in the MFRC context, is now defined by a linear truncated function $f(\cdot)$, and it monotonically increases with the object classifier response R_d as

$$f(\mathbf{x}_d, R_d) = \omega_d |\mathbf{x}_d| \max(0, R_d - R_t) \ , \tag{4}$$

where R_t is a threshold and ω_d defines the detector potential weight. Our region consistency constraint is similar to the object detector term used in [22] for scene understanding. If a detector response is strong, the higher-order potential will encourage the pixels belonging to the bounding box $\mathbf{x}_d$ to take the label l_d. As the penalty is increased with the number of inconsistent pixels incrementally until the truncation threshold Q_d, this soft higher-order constraint produces better labeling results than the standard P^n Potts model [17], which forbids other differently labeled pixels within the clique $\mathbf{x}_d$. The proposed object potential $\psi_d(\mathbf{x}_d)$ can be transformed to take the Robust P^n form [18, 22]:

$$\psi_d(\mathbf{x}_d) = -f(\mathbf{x}_d, R_d) + \min(f(\mathbf{x}_d, R_d), k_d \sum_{i \in \mathbf{x}_d} \delta(x_i \neq l_d)) \ , \tag{5}$$

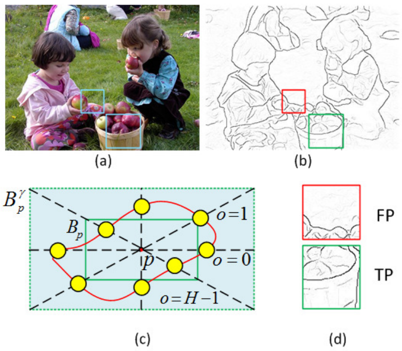

Figure 4. Proposed contour-layout filter based on the gPb contour map. (a) Input image. The blue bounding boxes indicate example detection results for the *apple bucket* image. (b) The intensity-inverted gPb map. (c) Proposed contour-layout filter (see the text for the details). (d) Close-up views of false/true positive object detections.

where k_d is a slope parameter defined in the same way as in [22]. Including this term to a CRF model is implemented by adding two auxiliary nodes into the graph, and the augmented energy function can be efficiently minimized with the graph cut algorithms. Interested readers are referred to [18] for the graph optimization details. Fig. 3 demonstrates the strength of the object detector-based potential when integrated into our CRF framework. Without using the detector-based potential, the black dog can only be partly annotated and segmented due to the weak low-level hypotheses based on pixels and segments. The object detector potential provides complementary high-level evidence and integrating it into the CRF model results in a more accurate result of recognizing the missed dog parts.

3.3. Contour for Object Boundary Reasoning

According to the cognition study [12], human vision views object part transitions at those with negative minima of curvature and the part salience depends on three factors: the relative size, the boundary strength and the degree of protrusion, so part transitions convey some mid-level information to help differentiate "object" and "stuff". Conventional contour detectors capture part transitions by finding local extrema, which usually result in a high recall but low precision contour detection result. Recently, Arbelazes *et al*. [4] proposed to combine the contour signal from eigenvectors with the low-level contour signal and achieved the state-of-the-art contour detection results. The eigenvectors v are generated by solving an eigen-system $(Z - W)v = \lambda Zv$, where $W = \{w_{ij}\}$ is a sparse symmetric affinity matrix encoding the pairwise similarity between image elements i and j based on the intervening contour cues [24, 4]. The diagonal matrix $Z = [z_{ij}]$ is defined with $z_{ii} = \sum_j w_{ij}$. As the affinity matrix W captures the global image information and the eigenvectors of the eigen-system are the solution to minimize the Ncut criteria [31], the eigenvectors capture the contour belonging to the transition between large object parts. This nice property makes gPb valuable for higher level image analysis. Fig. 4(b) shows an example gPb contour map C.

Contour-layout filters to reject false objects. Since our object detectors presented earlier use only color features for the robustness reason, the detected object proposals would unavoidably contain false positive detection results that belong to "stuff" such as sky. Inspired by the aforementioned part salience theory, we propose to exploit the gPb signal to define a novel objectness measure, which we call *contour-layout filters*. The basic idea is that if a detected bounding box falls on a non-object region, the contour distribution around the region tends to quite monotone, so we can reject this kind of detection results with high confidence. To extract such a distribution, we first enlarge a detected bounding box B_p centered at pixel p by a ratio γ, while preserving the original aspect ratio of B_p. The resulting relaxed rectangle B_p^γ defines the out-of-box bound. Next, we quantize the region around pixel p into H directions, and shoot H rays distributed evenly apart in angle (i.e. $2\pi/H$) from the center pixel p. If the ray for a quantized orientation bin $o \in \{0, 1, ..., H-1\}$ hits a salient gPb signal (with a strength above a threshold τ_{gPb}) within B_p^γ, we assign a value of 1 to the corresponding o-th component of a vector $V_p = [v_0, v_1, ..., v_{H-1}]^T$, otherwise 0 is assigned. In this paper we set $H = 8$, as shown in Fig. 4(c). To be robust to the spatial and orientation sampling discretization, we consider the contributions of the gPb responses of neighboring pixels in a small circular patch around the ray sampling location. Given this vector V_p, our contour-layout filters finally classify the detected object bounding box B_p as a false positive result, if the L1 norm of the vector V_p is less than an empirically predefined threshold $\tau_{cl} = 0.4$ for all classes. As the detector based potential is designed as a soft constraint in the inference, our method is not sensitive to the threshold and works well. We find this simple scheme is very effective in rejecting false object detections (see Fig. 4(d)) and preventing them from confusing the CRF inference, though more sophisticated methods to compute the objectness measure using the gPb contour can also be employed.

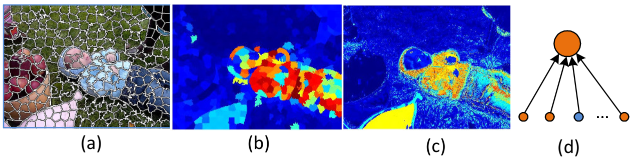

Figure 5. SLIC segmentation and segment-level consistency potentials. (a) SLIC segments [1]. (b) Segment-level and (c) pixel-level classifier response map for the *baby* foreground. (d) Robust region consistency constraint. The big orange dot denotes a segment assigned a label l_s. The blue dot denotes an outlier pixel.

Contour-based pairwise potential. Observing that the gPb contour map provides more reliable and higher-level reasoning of salient contours, we propose to compute the pairwise potential $\psi_{ij}(x_i, x_j)$ as follows,

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ \omega_a(1 - \|\nabla C(i,j)\|^2) & \text{otherwise ,} \end{cases} \quad (6)$$

where ω_a gives the weight of the pairwise potential. $\nabla C(i, j)$ measures the gPb signal contrast between two adjacent pixels i and j. We observe this new pairwise term reduces the possibility of incorrect boundary alignments compared with average color contrast based pairwise term [28].

4. Segment-Based Label Consistency Potential

The pixel-level features are usually too local to capture the change of neighborhood patterns, so we include an additional level of variables which consist of super-nodes/segments. We choose the SLIC algorithm [1] to over-segment the image into homogeneous regions. SLIC segments have been showed to give superior performance in terms of boundary adherence and segmentation compactness. Based on the generated super-nodes, we train a segment-based random forest classifier $\mathcal{C}_{seg}$. The feature computed at the segment level is the histogram of the textons with a dimension of T_c, which is defined earlier when producing the pixel-level features.

Now we present the formulation of the super-node based higher-order terms. Basically, we follow Pushmeet *et al.*'s method [18] to build a multi-layer hierarchical CRF model (two layers in our case), where the base layer consists of pixels and the second layer is made up of super-nodes which encode mid-level region cues. Such a construction enforces a soft constraint on the pixels belonging to a segment, encouraging them to be labeled as the same as their parent, but it also allows some outlier pixels (see Fig. 5(d)). Using a soft constraint makes our approach robust to the super-node quantization artifacts, while leveraging segments' grouping power and complementary cues extraction from a higher level for the given image. We have also tested hierarchical CRF models with more levels of super-nodes, and found that the results obtained are similar but at more computational costs. Fig. 5 shows a visual comparison between the segment-level classifier response map (color-coded as a heat map) and its pixel-level counterpart. One can notice that the segment-level classifier response is often stronger and more reliable than the pixel-level response, which tends to be noisy though with more details.

Given a segment $s \in S$ and its associated clique of pixels $\mathbf{x}_s$, let $N_s = \sum_{i \in \mathbf{x}_s} \delta(x_i \neq l_s)$ denote the number of variables in $\mathbf{x}_s$ not taking the segment label l_s, the super-node potential is designed as a linear truncated function [18]:

$$\psi_s(\mathbf{x}_s) = \omega_s . \begin{cases} N_s \frac{1}{Q_s}(\rho_{max} - \rho_{l_s}) + \rho_{l_s} & \text{if } N_s \leq Q_s \\ \rho_{max} & \text{otherwise ,} \end{cases} \qquad (7)$$

where ω_s is the weighting factor for the super-node potential. $\rho_{l_s} = -\log P(\mathbf{x}_s|\mathcal{C}_{seg})$ indicates the cost charge for a super-node $\mathbf{x}_s$ to take the label l_s. $P(\mathbf{x}_s|\mathcal{C}_{seg})$ is given by a random forest super-node classifier $\mathcal{C}_{seg}$. ρ_{max} is the maximum cost charge when a number of Q_s pixels do not take the label l_s. This segment potential can also be finally transformed to the Robust P^n form:

$$\psi_s(\mathbf{x}_s) = \min\{\min_{l_s}(\rho_{l_s} + k_{l_s} \sum_{i \in \mathbf{x}_s} \delta(x_i \neq l_s)), \rho_{max}\} \ , \qquad (8)$$

where k_{l_s} is the slope parameter similarly defined as k_d in (5). Similar to the object detector potential, this term can be minimized by including two auxiliary node in the graph and solved efficiently with graph cut [18].

5. Experimental Results and Discussions

We evaluate our method using the FlickrMFC dataset [15]. This dataset is the **ONLY** MFC dataset consists of 14 groups of images with manually labeled ground-truth. Each group includes **10∼20** images which are sampled from a Flickr photostream. This dataset is challenging as it contains a finite number of repeating subjects that are not presented in every image and there are strong lighting variation, pose change and background clutters in the images. The parameters are empirically fixed as: $\omega_{pix} = 1, \omega_a = 10, \omega_s = 0.2, \omega_d = 0.1, R_t = 0.5, \rho_{max} = -\log(0.1)$. The overall time (including preprocessing, detection and segmentation) to process each image is around 20∼30 seconds on a desktop Intel Core i5 3.2GHZ and 8GB RAM.

Quantitative Results: We compare our method with some baselines: MFC [15], CoSand(COS) [16], Discriminative Clustering(DC) [14], LDA [30]. We adopt the procedure introduced in MFC [15] for evaluation. For supervised methods such as our method and MFC's supervised version, we randomly pick 20% of the input images to annotate. For the unsupervised methods, e.g. CoSand [16], DC [14] and LDA [30], the dataset is divided into several subgroups such that the images in each subgroup contain the same objects of interest, the methods are applied to each subgroup individually. We evaluate the segmentation accuracy by the standard intersection-over-union metric $\frac{(GT_i \cap R_i)}{(GT_i \cup R_i)}$.

Fig. 6 summarizes the segmentation accuracy on the 14 groups of the FlickrMFC dataset. The left most bar set presents the average segmentation accuracy on 14 groups. Since COS, DC, LDA and MFC-U are unsupervised methods which count on low-level cues, they failed to capture the real objects of interest, so their performance is not competitive in most cases. We hence focus on the comparison with the state-of-the-art MFC method. As shown in the bar chart, our algorithm's average accuracy is around 10% higher than the MFC method [15]. Some datasets like *cheetah*, *butterfly*, *liberty*, we achieve around 20% accuracy improvement. For the *thinker* dataset, the accuracy gap reaches even 50%!

We have also evaluated the average accuracy gain contributed by including higher-order segment and detector potentials into the CRF model, which is about 2.3%. This numerical small gain has also been observed in Shotton *et al.* [32] and Pushmeet *et al.* [18] in scene understanding research. As also indicated in [32, 18], we observe that including these potentials often bring a pronounced increase in perceived accuracy, especially for the challenging cases such as Fig. 1, 3 and 8.

Visual Results: Fig. 7 shows some visual results from seven groups of FlickrMFC dataset. For each set, the input images and color-coded segmentation results are displayed in the first two rows from top to bottom. The regions which are labeled with the same color in each set indicate they belong to the same category. The tags below each set explain the meaning of each color. From the images, one can observe that our method can handle irregularly appearing objects and produce smooth and accurate segmentation results. The images with no foregrounds are also correctly identified, e.g the *liberty* dataset. On the other hand, our current model still cannot handle some camouflage cases very well, such as the *butterfly* dataset.

6. Conclusion and Future Work

We proposed the MFRC framework which performs joint detection, recognition and segmentation of multiple foreground objects that irregularly occur in each image. To further improve the detection accuracy of our method, exploiting some structure information at the object level can be helpful. Moreover, how to extend

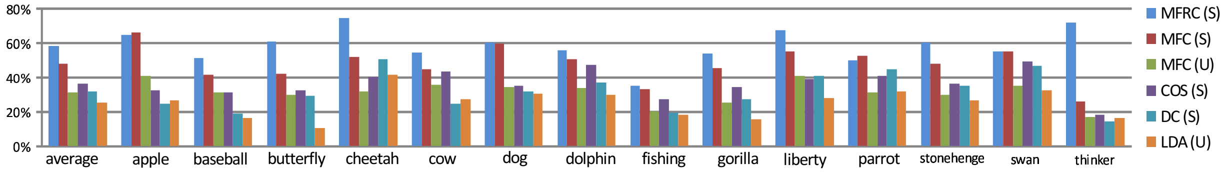

Figure 6. Segmentation accuracy comparison between our method (MFRC) and other baselines (MFC [15], COS [16], DC [14], LDA [30]) for the FlickrMFC dataset. The S and U denote whether the method is supervised or unsupervised.

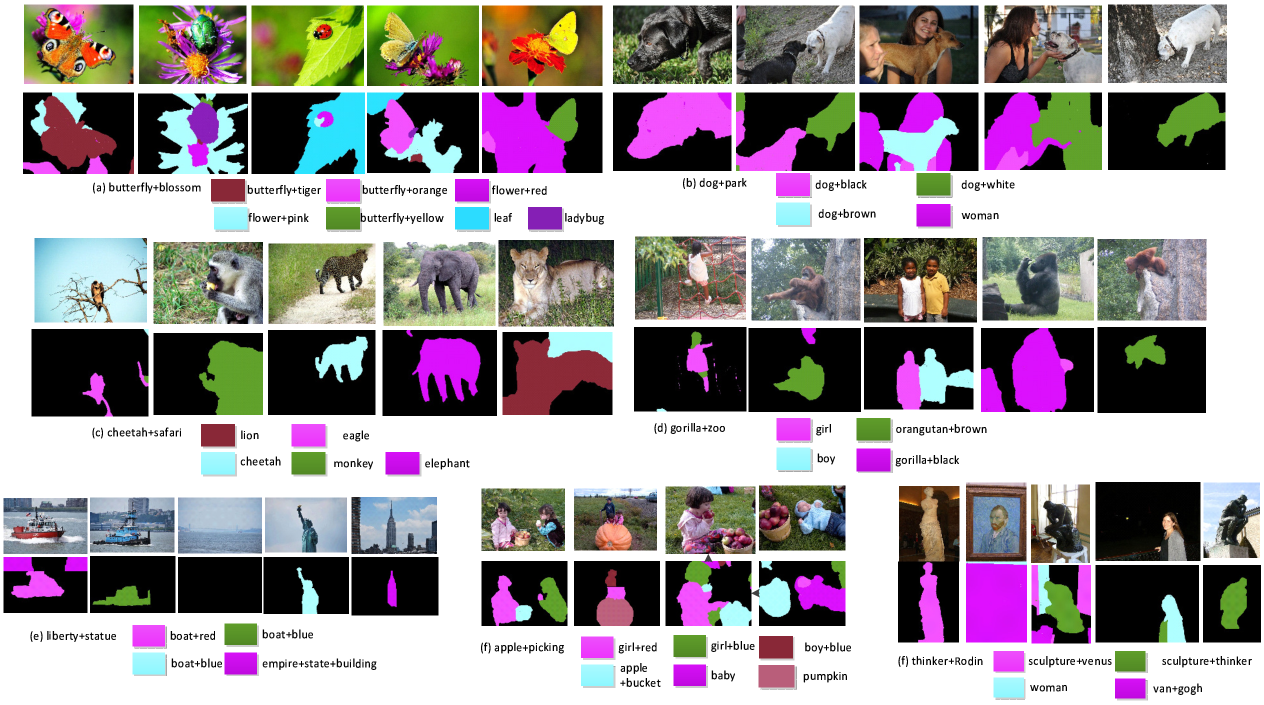

Figure 7. Some randomly drawn examples from seven groups of the FlickrMFC dataset. From top to bottom, each set presents its input images, color-labeled segmentation results. The colored tag below each set indicates which category each region is assigned to.

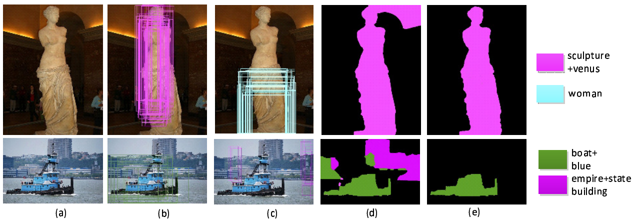

Figure 8. Behavior of the object detector potential as a soft constraint. (a) Input image. (b,c) Two object detection results. Our results (d) without and (e) with using the object detector potentials.

our algorithm to a large scale image dataset is an interesting topic for future work.

7. Acknowledgement

This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR). The authors would also like to acknowledge the Ph.D. grant from the IMI, NTU, Singapore.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11), 2012. 5, 6

[2] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010. 2

[3] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7), 1997. 3

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI.*, 33(5), 2011. 3, 5

[5] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. Interactively co-segmentating topically related images with intelligent scribble guidance. *International Journal of Computer Vision*, 93(3), 2011. 2

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11), 2001. 2

[7] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *IEEE TPAMI.*, 34(7), 2012. 2

[8] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *CVPR*, 2011. 1, 2

[9] M. D. Collins, J. Xu, L. Grady, and V. Singh. Random walks based multi-image segmentation: Quasiconvexity results and gpu-based solutions. In *CVPR*, 2012. 2

[10] P. F. Felzenszwalb, R. B. Girshick, and D. A. McAllester. Cascade object detection with deformable part models. In *CVPR*, 2010. 2, 4

[11] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, 71(3), 2007. 2, 4

[12] D. D. Hoffman and M. Singh. Salience of visual parts. *Cognition*, 63(1):29 – 78, 1997. 5

[13] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 1, 2

[14] A. Joulin, F. R. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010. 1, 2, 6, 7

[15] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *CVPR*, pages 837–844, 2012. 1, 2, 3, 6, 7

[16] G. Kim, E. P. Xing, F.-F. Li, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 1, 2, 6, 7

[17] P. Kohli, M. P. Kumar, and P. H. S. Torr. P & beyond: Move making algorithms for solving higher order functions. *IEEE TPAMI.*, 31(9), 2009. 4

[18] P. Kohli, L. Ladicky, and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3), 2009. 2, 4, 5, 6

[19] D. Küttel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR*, 2012. 2

[20] D. Küttel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In *ECCV*, 2012. 2

[21] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010. 2, 4

[22] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, 2010. 2, 4, 5

[23] T. Ma and L. J. Latecki. Graph transduction learning with connectivity constraints with application to multiple foreground cosegmentation. In *CVPR*, 2013. 2

[24] D. R. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE TPAMI.*, 2004. 5

[25] L. Mukherjee, V. Singh, and C. R. Dyer. Half-integrality based algorithms for cosegmentation of images. In *CVPR*, 2009. 1

[26] L. Mukherjee, V. Singh, and J. Peng. Scale invariant cosegmentation for image groups. In *CVPR*, 2011. 1, 2

[27] F. M. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *CVPR*, 2005. 4

[28] C. Rother, V. Kolmogorov, and A. Blake. ”grabcut”: interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3), 2004. 3, 5

[29] J. C. Rubio, J. Serrat, A. M. López, and N. Paragios. Unsupervised co-segmentation through region matching. In *CVPR*, 2012. 1, 2

[30] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 6, 7

[31] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000. 5

[32] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1), 2009. 2, 4, 6

[33] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009. 4

[34] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, 2011. 2

[35] Y. Wei, J. Sun, X. Tang, and H.-Y. Shum. Interactive offline tracking for color objects. In *ICCV*, 2007. 2, 4